handbook of statistics 30

Time Series Analysis: Methods and Applications

Edited by
T. Subba Rao
S. Subba Rao
C.R. Rao

HANDBOOK OF STATISTICS
VOLUME 30

Handbook of Statistics

VOLUME 30

General Editor

C.R. Rao
C.R. Rao AIMSCS, University of Hyderabad Campus, Hyderabad, India

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Volume 30

Time Series Analysis: Methods and Applications

Edited by

Tata Subba Rao
University of Manchester, UK

Suhasini Subba Rao
Texas A&M University, College Station, Texas, USA

C.R. Rao
C.R. Rao AIMSCS, University of Hyderabad Campus, Hyderabad, India

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

North-Holland is an imprint of Elsevier

North-Holland is an imprint of Elsevier
The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, UK
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands

First edition 2012

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress.

ISBN: 978-0-444-53858-1
ISSN: 0169-7161

For information on all North-Holland publications visit our web site at *store.elsevier.com*

Typeset by: diacriTech, India

Printed and bound in Great Britain
12 13 14 15 10 9 8 7 6 5 4 3 2 1

Table of Contents

Volume 30 Time Series

Preface to Handbook Volume – 30

Nearly 25 years ago, two volumes on time series (volume 3 in 1983 and volume 5 in 1985) were published as a part of the Handbook of Statistics series. Volume 3 is on "Time Series in the Frequency domain" edited by D. R. Brillinger and P. R. Krishnaiah, which contains papers related to spectral methods of time series (mainly linear times and to a lesser extent nonlinear time series). Volume 5 is on "Time Series in the Time Domain" edited by E. J. Hannan, P. R. Krishnaiah, and M. M. Rao, which comprises of papers on statistical inference of linear ARMA models, nonstationary spectral representations, various order selection procedures, etc.

Since the above publications, there has been an explosion of developments in time series, for example, Bootstrap methods for time-dependent data, nonlinear models for time series, analysis of high-frequency time series, quantile regression, time and frequency domain methods for the analysis of data from biological, environmental sciences, and so on. The main aim of the present volume is to survey these recent developments. This volume is divided into 10 parts, covering some of the areas mentioned above, our object here is to make each part as homogeneous as possible.

It seems appropriate to start the volume by covering topics that were in their infancy 25 years ago. Part I of the volume covers bootstrap methods and tests for linearity of a time series. Kreiss and Lahiri review bootstrap methods for dependent data satisfying various linear time series models, nonlinear models and processes with long memory properties. In Chapter 2, Berg, McMurry, and Politis consider some bootstrap methods for studying the asymptotic properties of linearity tests based on the higher order spectra. There is further discussion on tests for linearity in Chapter 3 by Giannerini.

In the case that we reject the null in a test for linearity, it then becomes necessary to consider methods of modeling nonlinear time series, and this is the focus of Part II of the volume. In Chapter 4, Tjøstheim gives an overview of modeling methods for nonlinear, nonstationary time series. Two important examples of nonlinear times series commonly used in financial time series are Markov switching models and ARCH-type models, which are considered in the next two chapters. In Chapter 5, Franke considers both the probabilistic properties of various Markov switching models and also inference of such models, including estimation and model selection. In Chapter 6, Mukherjee considers robust estimation of parameters for heteroscedastic models including ARCH and GARCH models.

Part III consists of papers related to functional data and high-dimensional time series, two areas that are currently receiving a lot of attention in the wider statistical community. Hörmann and Kokoszka, in Chapter 7, provide a functional time series

approach to periodically stationary continuous time series. In Chapter 8, Wu and Xiao consider the estimation of covariance matrices of both low- and high-dimensional time series and study their asymptotic sampling properties under the assumption of physical dependence.

Part IV comprises of just one paper by Zijie Xiao, in which he gives an overview of time series quantile regression methods. He starts by reviewing conditional quantiles for classical time series models, for example, linear and nonlinear time series and discusses how quantiles can be used to expand on current time series models, for example, quantile autoregressive models.

Part V contains papers where applications to biological and neurological sciences are considered (both time and frequency domain approaches are used). In Chapter 10, Stoffer gives an overview of the spectral envelope for the harmonic analysis and scaling of categorical-valued time series and demonstrates how this methodology can be used to search for diagnostic patterns in DNA sequences. Ozaki, in Chapter 11, considers fMRI data (functional magnetic resonance imaging) using spatial time series models and the Kalman-Bucy filtering algorithm. Dependent count data arise in many biological applications, and in Chapter 12, Fokianos considers various time series models for such data, including both Poisson regression models and integer-valued processes.

Part VI covers nonstationary time series. The main focus will be on time series whose structure evolves slowly over time and are considered "locally stationary processes". A term that is precisely defined in the comprehensive overview on locally stationary processes, for both linear and nonlinear time series, given by Dahlhaus in Chapter 13. The following two papers consider representations of a nonstationary time series. Ombao, in Chapter 14, starts by defining the windowed Fourier basis, which is a flexible basis for time–frequency localization of a signal. He then uses this basis to search for a parsimonious representation of the time-varying spectral density function. These methods are applied to the analysis of electroencephalograms (EEGs) data. In Chapter 15, Subba Rao considers the stochastic coefficient regression (SCR) model, a variant of the classical multiple regression model that takes into account the influence regressors may have on the variance of the response variable. She shows that locally stationary processes can be represented as a SCR model and considers inference based on the Fourier transform of the observations.

Part VII contains papers on spatio-temporal models. Although there exists a vast literature on spatial processes, only recently has more attention been given to spatio-temporal processes, in view of the extra temporal dimension. In Chapter 16, Sahu discusses a hierarchical autoregressive Bayesian model for space–time air pollution data and also considers ways of monitoring ozone pollution. Fontanella and Ippoliti, in Chapter 17, discuss the Karhunen–Loeve expansion of spatio-temporal processes and consider how this expansion can be applied in the climatological series. In Chapter 18, Subba Rao and Terdik start by reviewing the literature on the statistical analysis of spatial processes and then define measures of dependence and propose alternative methods based on the discrete Fourier transform for the estimation of spatio-temporal models.

All the previous chapters have addressed issues related to discrete time series, however continuous time series also arise in several applications and this will be the focus of part VIII. An important application of continuous time series is in financial time series, where until recently most continuous time series models considered were driven

by a Wiener process that restricts the sample paths to be continuous. This can be a restrictive assumption, therefore recently the more general Levy-driven continuous time series models have received much attention. These processes are reviewed by Brockwell and Lindner, who consider their theoretical properties and show how they can be used to model financial time series. Many discrete time series are samples from a continuous-time process, therefore it is important to understand how the two time series are related. In Chapter 20, Turkman considers the relationship between the supremum of a stationary continuous time and the maximum over the discretized version of the time series.

Part IX considers spectral and wavelet methods for the analysis of signals. Quinn considers the Fourier analysis of time series for the estimation of frequencies in a signal in Chapter 21. In Chapter 22, Percival and Mondall consider wavelet methods for the analysis of variance in time series.

A sign of the times is that many of the methodologies described above are easily available as software packages written in `R` (the statistical software used by many statisticians). Therefore, it seems timely to include a chapter on `R`. Part X comprises of one chapter by McLeod, Yu, and Mahdi, who give an introduction to time series analysis in `R`.

In editing this volume, we received help from the contributors who voluntarily reviewed the papers (sometimes more than once), and we wish to thank all of them. We are also grateful to Professor C. W. Anderson, Sheffield University, Professor M. B. Priestley, University of Manchester, and Professor M. Pourahamadi, Texas A&M University who reviewed some of the papers.

T. Subba Rao
S. Subba Rao
C. R. Rao

This page intentionally left blank

Contributors: Vol. 30

Berg, Arthur, *Division of Biostatistics, Penn State University, Hershey, PA 17033, USA; e-mail: berg@psu.edu* (Ch. 2).

Brockwell, Peter, *Department of Statistics, Colorado University, Fort Collins, Colorado 80523-1877, USA; e-mail: peter.brockwell@gmail.com* (Ch. 19).

Dahlhaus, Rainer, *Institut für Angewandte Mathematik, Universität Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany; e-mail: dahlhaus@statlab.uni-heidelberg.de* (Ch. 13).

Fokianos, Konstantinos, *Department of Mathematics & Statistics, University of Cyprus, Nicosia 1678, Cyprus; e-mail: fokianos@ucy.ac.cy* (Ch. 12).

Fontanella, Lara, *Department of Quantitative Methods and Economic Theory, University G. d'Annunzio, Viale Pindaro 42, 65127 Pescara, Italy; e-mail: lfontan@dmqte.unich.it* (Ch. 17).

Franke, Jürgen, *Department of Mathematics, University of Kaiserslautern, Erwin-Schroedinger-Str., 67663 Kaiserslautern, Germany; e-mail: franke@mathematik.uni-kl.de* (Ch. 5).

Giannerini, Simone, *Dipartimento di Scienze Statistiche, Università di Bologna, Via Belle Arti 41, 40126 Bologna, Italy; e-mail: simone.giannerini@unibo.it* (Ch. 3).

Hörmann, Siegfried, *Department of Mathematics, Université Libre de Bruxelles, Bd du Triomphe, B-1050 Bruxelles, Belgique; e-mail: siegfried.horman@ulb.ac.be* (Ch. 7).

Ippoliti, Luigi, *Department of Quantitative Methods and Economic Theory, University G. d'Annunzio, Viale Pindaro 42, 65127 Pescara, Italy; e-mail: ippoliti@unich.it* (Ch. 17).

Kokoszka, Piotr, *Department of Statistics, Colorado State University, Fort Collins, Colorado 80523-1877, USA; e-mail: piotr.kokoszka@colostate.edu* (Ch. 7).

Kreiss, Jens-Peter, *Technische Universität Braunschweig, Institut für Mathematische Stochastik, Pockelsstrasse 14, D-38106 Braunschweig, Germany; e-mail: j.kreiss@tu-bs.de* (Ch. 1).

Lahiri, Soumendra Nath, *Department of Statistics, TAMU-3143 Texas A & M University, College Station, TX 77843, USA; e-mail: snlahiri@stat.tamu.edu* (Ch. 1).

Lindner, Alexander, *Institut für Mathematische Stochastik TU Braunschweig, Germany; e-mail: a.lindner@tu-bs.de* (Ch. 19).

Mahdi, Esam, *Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ontario, Canada N6A 5B7; e-mail: emahdi@statsuwo.ca* (Ch. 23).

McLeod, A. Ian, *Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ontario, Canada N6A 5B7; e-mail: aim@stats.uwo.ca* (Ch. 23).

McMurry, Timothy, *Division of Biostatistics, University of Virginia, Charlottesville, VA 22908-0717, USA; e-mail: tlm6w@virginia.edu* (Ch. 2).

Mondal, Debashis, *Department of Statistics, University of Chicago, South University Avenue, Chicago, IL, USA; e-mail: mondal@galton.uchicago.edu* (Ch. 22).

Mukherjee, Kanchan, *Department of Mathematics and Statistics, Lancaster University, Lancaster LA14YF, United Kingdom; e-mail: k.mukherjee@lancaster.ac.uk* (Ch. 6).

Ombao, Hernando, *Center of Statistical Science, Brown University, 121 South Main street, 7th Floor, Providence, RI 02912, USA; e-mail: ombao@stat.brown.edu* (Ch. 14).

Ozaki, Tohru, *IDAC, Tohoku University, Sendai, Japan; e-mail: tohru.ozaki@gmail.com* (Ch. 11).

Percival, Donald B., *Applied Physics Laboratory, University of Washington, Seattle, WA, USA; e-mail: dbp@apl.washington.edu* (Ch. 22).

Politis, Dimitris N., *Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112, USA; e-mail: dpolitis@ucsd.edu* (Ch. 2).

Quinn, Barry G., *Department of Statistics, Maquaire University, NSW 2109, Australia; e-mail: bquinn@efs.mq.edu.au* (Ch. 21).

Sahu, Sujit K., *School of Mathematics, University of Southampton, Southampton SO17 1BJ, United Kingdom; e-mail: S.K.Sahu@soton.ac.uk* (Ch. 16).

Stoffer, David S., *Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, USA; e-mail: stoffer@pitt.edu* (Ch. 10).

Subba Rao, Suhasini, *Department of Statistics, Texas A & M University, College Station, TX 77843, USA; e-mail: suhasini@stat.tamu.edu* (Ch. 15).

Subba Rao, Tata, *School of Mathematics, The University of Manchester, Manchester M13 9Pl, United Kingdom; C. R. RAO AIMSCS University of Hyderabad campus, Hyderabad, India; e-mail: tata.subbarao@manchester.ac.uk* (Ch. 18).

Terdik, Gyorgy, *Faculty of Informatics, University of Debrecen, Debrecen, Hungary; e-mail: terdik@delfin.unideb.hu* (Ch. 18).

Tjøstheim, Dag, *Department of Mathematics, University of Bergen, Johs. Brunsgt. 12, 5008 Bergen, Norway; e-mail: dag.tjøstheim@math.uib.no* (Ch. 4).

Turkman, K. F., *Departamento de Estatstica e Investigao Operacional, Universidade de Lisboa, Lisboa 1749-016, Portugal; e-mail: kfturkman@fc.ul.pt* (Ch. 20).

Wu, Wei Biao, *Department of Statistics, The University of Chicago, Chicago, IL 60637, USA; e-mail: wbwu@galton.uchicago.edu* (Ch. 8).

Xiao, Han, *Department of Statistics, The University of Chicago, Chicago, IL 60637, USA; e-mail: xiao@aitken-cluster.uchicago.edu* (Ch. 8).

Xiao, Zhijie, *Department of Economics, Boston College, Chestnut Hill, MA 02467, USA; e-mail: zhijie.Xiao@bc.edu* (Ch. 9).

Yu, Hao, *Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ontario, Canada N6A 5B7; e-mail: hyu@stats.uwo.ca* (Ch. 23).

Part I: Bootstrap and Tests for Linearity of a Time Series

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00001-6

Bootstrap Methods for Time Series

Jens-Peter Kreiss[1] *and Soumendra Nath Lahiri*[2]

[1]*Technische Universität Braunschweig, Institut für Mathematische Stochastik, Pockelsstrasse 14, D-38106 Braunschweig, Germany*
[2]*Department of Statistics, TAMU-3143 Texas A & M University, College Station, TX 77843, USA*

Abstract

The chapter gives a review of the literature on bootstrap methods for time series data. It describes various possibilities on how the bootstrap method, initially introduced for independent random variables, can be extended to a wide range of dependent variables in discrete time, including parametric or nonparametric time series models, autoregressive and Markov processes, long range dependent time series and nonlinear time series, among others. Relevant bootstrap approaches, namely the intuitive residual bootstrap and Markovian bootstrap methods, the prominent block bootstrap methods as well as frequency domain resampling procedures, are described.

Further, conditions for consistent approximations of distributions of parameters of interest by these methods are presented. The presentation is deliberately kept non-technical in order to allow for an easy understanding of the topic, indicating which bootstrap scheme is advantageous under a specific dependence situation and for a given class of parameters of interest. Moreover, the chapter contains an extensive list of relevant references for bootstrap methods for time series.

Keywords: bootstrap methods, discrete Fourier transform, linear and nonlinear time series, long range dependence, Markov chains, resampling, second order correctness, stochastic processes.

1. Introduction

The bootstrap method, initially introduced by Efron (1979) for independent variables and later extended to deal with more complex dependent variables by several authors, is a class of nonparametric methods that allow the statistician to carry out statistical

inference on a wide range of problems without imposing much structural assumptions on the underlying data-generating random process. By now, there exist several books and monographs, e.g., Hall (1992), Efron and Tibshirani (1993), Shao and Tu (1995), Davison and Hinkley (1997), and Lahiri (2003a), among others, which describe different aspects of the bootstrap methodology at varying levels of sophistication and generality. Moreover, several papers in the literature give overviews of various aspects of bootstrapping time series. Among them are Berkowitz and Kilian (2000), Bose and Politis (1995), Bühlmann (2002), Carey (2005), Härdle et al. (2003), Li and Maddala (1996), and Politis (2003). These papers consider bootstrap and resampling methods for general stochastic processes and time series models. The review papers by Paparoditis and Politis (2009) and by Ruiz and Pascual (2002) especially focus on financial time series, while McMurry and Politis considers resampling methodology for functional data. In this article, we aim to provide an easy-to-read description of some of the key ideas and issues and present latest results on a set of selected topics in the context of time series data showing temporal dependence.

The basic idea behind the bootstrap methods is very simple, and it can be described in general terms as follows. Let $X_1, \ldots, X_n$ be a stretch of a time series with joint distribution P_n. For estimating a population parameter θ, suppose that we have constructed an estimator $\hat{\theta}_n$ (e.g., using the generalized method of moments) based on $X_1, \ldots, X_n$. A common problem that the statistician must deal with is to assess the accuracy of $\hat{\theta}_n$, for example, by using an estimate of its mean squared error (MSE) or an interval estimate of a given confidence level. However, any such measure of accuracy depends on the sampling distribution of $\hat{\theta}_n - \theta$, which is typically unknown in practice and often very complicated. Bootstrap methods provide a general recipe for estimating the distribution of $\hat{\theta}_n$ and its functionals without restrictive model assumptions on the time series.

We now give a general description of the basic principle underlying the bootstrap methods. As before, suppose that the data are generated by a part of a time series $\{X_1, \ldots, X_n\} \equiv \mathbf{X}_n$ with joint distribution P_n. Given $\mathbf{X}_n$, first construct an estimate $\hat{P}_n$ of P_n. Next, generate random variables $\{X_1^*, \ldots, X_n^*\} \equiv \mathbf{X}_n^*$ from $\hat{P}_n$. If $\hat{P}_n$ is a reasonably "good" estimator of P_n, then the relation between $\{X_1, \ldots, X_n\}$ and P_n is closely reproduced (in the bootstrap world) by $\{X_1^*, \ldots, X_n^*\}$ and $\hat{P}_n$. Define the bootstrap version $\hat{\theta}_n^*$ of $\hat{\theta}_n$ by replacing $X_1, \ldots, X_n$ with $X_1^*, \ldots, X_n^*$, and similarly, define θ^* by replacing P_n in $\theta = \theta(P_n)$ by $\hat{P}_n$. Then, the conditional distribution (function) $\hat{G}_n$ or G_n^* (say) of $\hat{\theta}_n^* - \theta^*$ (given $\mathbf{X}_n$) gives the bootstrap estimator of the distribution (function) G_n (say) of $\hat{\theta}_n - \theta$. Here, θ^* is some properly chosen parameter, which in many applications can be computed from $\hat{P}_n$ along the same lines as θ is computed from P_n. In almost all applications, the bootstrap is used to approximate distributions of the type $c_n(\hat{\theta}_n - \theta)$, where the to infinity increasing sequence (c_n) of non-negative real numbers is chosen such that the sequence of distributions converges to a nondegenerate limit.

To define the bootstrap estimators of a functional of the distribution of $\hat{\theta}_n - \theta$, such as the variance or the quantiles of $\hat{\theta}_n - \theta$, we may simply use the "plug-in" principle and employ the corresponding functional to the conditional distribution of $\hat{\theta}_n^* - \theta^*$. Thus, the bootstrap estimator of the variance σ_n^2 of $\hat{\theta}_n - \theta$ is given by the conditional

variance $\hat{\sigma}_n^2$ of $\hat{\theta}_n^* - \theta^*$, i.e., by

$$\begin{aligned}\hat{\sigma}_n^2 &= \text{the bootstrap estimator of } \sigma_n^2 \\ &= \text{Var}(\hat{\theta}_n^* - \theta^* | \mathbf{X}_n) \\ &= \int x^2 d\hat{G}_n(x) - \left[\int x d\hat{G}_n(x)\right]^2.\end{aligned}$$

Similarly, if $q_{\alpha,n}$ denotes the $\alpha \in (0,1)$ quantile of (the distribution of) $\hat{\theta}_n - \theta$, then its bootstrap estimator is given by

$$\hat{q}_{\alpha,n} = \hat{G}_n^{-1}(\alpha), \text{ the } \alpha \text{ quantile of the conditional distribution of } \hat{\theta}_n^* - \theta^*.$$

In general, having chosen a particular bootstrap method for a specific application, it is very difficult (and often, impractical) to derive closed-form analytical expressions for the bootstrap estimators of various population quantities. This is where the computer plays an indispensable role. Bootstrap estimators of the distribution of $\hat{\theta}_n - \theta$ can be computed numerically using Monte-Carlo simulation. First, a large number (usually in hundreds) of independent copies $\{\hat{\theta}_n^{*k} : k = 1, \ldots, K\}$ of $\hat{\theta}_n^*$ are constructed by *repeated* resampling. The empirical distribution of these bootstrap replicates gives the desired Monte-Carlo approximation to the true bootstrap distribution of $\hat{\theta}_n^* - \theta^*$ and to its functionals. Specifically, for the variance parameter $\sigma_n^2 = \text{Var}(\hat{\theta}_n - \theta)$, the Monte-Carlo approximation to the bootstrap estimator $\hat{\sigma}_n^2$ is given by

$$[\hat{\sigma}_n^{\text{MC}}]^2 \equiv (K-1)^{-1} \sum_{k=1}^{K} \left[\hat{\theta}_n^{*k} - K^{-1} \sum_{j=1}^{K} \hat{\theta}_n^{*j}\right]^2,$$

the sample variance of the replicates $\{\hat{\theta}_n^{*k} - \theta^* : k = 1, \ldots, K\}$. Similarly, the Monte-Carlo approximation to the bootstrap estimator $\hat{q}_{\alpha,n}$ is given by

$$\hat{q}_{n,\alpha}^{\text{MC}} \equiv \hat{\theta}_n^{*(\lfloor K\alpha \rfloor)} - \theta^*,$$

the $\lfloor K\alpha \rfloor$ order statistic of the replicates $\{\hat{\theta}_n^{*k} - \theta^* : k = 1, \ldots, K\}$, where for any real number x, $\lfloor x \rfloor$ denotes the largest integer not exceeding x. From this point of view, the introduction of the bootstrap has been very timely; almost none of the interesting applications of the bootstrap would have been possible without the computing power of present day computers.

The rest of the paper is organized as follows. Section 2 presents and discusses residual bootstrap methods for parametric and nonparametric models. The proposals mainly apply the classical bootstrap approach of *drawing with replacement* to residuals of a fitted model to the data. As a special case, Section 3 considers in detail an approach by fitting autoregressions of increasing order to the observed data. A rather relevant model class of dependent observations to which bootstrap procedures successfully can be applied are Markov chains (cf. Section 4).

Section 5 discusses in detail the prominent block bootstrap methods for time series. So far, all discussed bootstrap methods are in time domain. Of course, frequency domain bootstrap methods exist and are presented in Section 6. Mixtures of both frequency and time domain bootstrap methods are described in Section 7. A final Section 8 concentrates on bootstrap methods for time series with long-range dependence.

2. Residual bootstrap for parametric and nonparametric models

Since the original bootstrap idea of Efron (1979) for i.i.d. random variables of *drawing with replacement* cannot be applied directly to dependent observations, because by obvious reasons, it suggests itself to apply the classical bootstrap principle to residuals of an (optimal) predictor of the $X_t's$.

Suppose for the following that we are given observations $X_1, \ldots, X_n$. For some fixed $p \in \mathbb{N}$ denote by $\widehat{m}_n(X_{t-1}, \ldots, X_{t-p})$, a parametric or nonparametric estimator of the conditional expectation $E[X_t | X_{t-1}, \ldots, X_{t-p}]$. This estimator leads to residuals

$$\widehat{e}_t := X_t - \widehat{m}_n(X_{t-1}, \ldots, X_{t-p}), t = p+1, \ldots, n, \tag{1}$$

and in a next step to a bootstrap time series

$$X_t^* = \widehat{m}_n(X_{t-1}^*, \ldots, X_{t-p}^*) + e_t^*, t = 1, \ldots, n. \tag{2}$$

The bootstrap innovations $e_1^*, \ldots, e_n^*$ follow a Laplace distribution over the set $\{\widehat{e}_{p+1}^c, \ldots, \widehat{e}_n^c\}$ of centered estimated residuals $\widehat{e}_{p+1}, \ldots, \widehat{e}_n$.

Here, we presumed that all residuals more or less share the same variance. In a heteroscedastic situation, one might think of some kind of a localized selection of bootstrap residuals or a *wild* bootstrap approach. The latter means that bootstrap innovations are generated according to

$$e_t^* := \widehat{e}_t \cdot \eta_t^*, t = p+1, \ldots, n, \tag{3}$$

where the (bootstrap) random variables (η_t^*) possess zero mean and unit variance, only. Typically, it is not necessary to specify some distribution for the η_t^*'s. If a distributional assumption is made, this ranges from rather simple discrete (even two-point) distributions to standard normal distribution. For reasons of better higher order performance for properly studentized statistics, one additionally should ensure $E^*(\eta_t^*)^3 = 1$. The simple discrete distribution taking values $z_1 = (1+\sqrt{5})/2$ and $z_2 = (1-\sqrt{5})/2$ with probabilities $p_1 = (\sqrt{5}-1)/(2\sqrt{5})$ and $p_2 = (\sqrt{5}+1)/(2\sqrt{5})$, respectively, satisfies the assumption of zero mean and unit second and third moments.

If we decide to use a fully nonparametric estimator in (1), the probabilistic properties of the bootstrap time series (2) could be rather delicate to investigate, because we, in principle, could not control the behavior of nonparametric estimators in regions far away from the origin, because we do not have many underlying observations in such regions. This typically leads to not very reliable estimators in that regions, and therefore, the stability of the bootstrap process cannot easily be guaranteed (recall that

models of type (2) typically need some quite restrictive growth conditions on the behavior of the function $\widehat{m}_n(x_{t-1}, \ldots, x_{t-p})$). But in order to establish asymptotic consistency of this bootstrap proposal, we need at least some stability and typically moreover some mixing or weak dependence properties for the triangular array of dependent observations in the bootstrap world. Such conditions would be rather helpful in order to prove asymptotic results for the bootstrap process.

One way out of this problem is to define instead of (2), a regression model in the bootstrap world, i.e., to generate bootstrap observations according to

$$X_t^* = \widehat{m}_n(X_{t-1}, \ldots, X_{t-p}) + e_t^*, t = 1, \ldots, n. \tag{4}$$

Along this proposal, we do not obtain a time series in the bootstrap world any longer, but an advantage of this proposal over (2) is that the design variables (which are lagged original observations themselves) now indeed mimic the p-dimensional marginal distribution of the underlying data by construction.

The investigation of a residual bootstrap procedure is much simpler; hence, we decide to use a fully parametric estimator in (1). For example, an optimal linear approximation of the conditional expectation, i.e., an autoregressive fit of order p to the underlying data. The estimator $\widehat{m}_n$ in this case simplifies to $\widehat{m}_n(x_1, \ldots, x_p) = \sum_{k=1}^{p} \widehat{a}_k x_{t-k}$. Using Yule-Walker parameter estimates $\widehat{a}_k$ in such a simple situation always leads to a stable and causal process in the bootstrap world (cf. Kreiss and Neuhaus (2006), Satz 8.7 and Bemerkung 8.8). But, of course, one can apply the idea of a parametric fit to the conditional expectation to other models including moving-average and ARMA models.

The question of main interest is in which situations and to what extent the described bootstrap proposals asymptotically work.

In order to ensure that a fitted *parametric* model generates according to (2) bootstrap data that are able to mimic all dependence properties of the underlying observations, one has to assume that the data-generating process itself belongs to the parametric class, i.e., possess a representation of the form

$$X_t = m_\theta(X_{t-1}, \ldots, X_{t-p}) + e_t, \, t \in \mathbb{Z}, \tag{5}$$

with i.i.d. innovations and parametric conditional mean function m_θ, which of course is quite restrictive. However, it can be stated that the parametric residual bootstrap consistently mimics the process (5). An obvious extension of the residual bootstrap (including an estimator of the conditional deviation (volatility)) leads to a residual bootstrap which consistently mimics the following slight deviation of model (5)

$$X_t = m_\theta(X_{t-1}, \ldots, X_{t-p}) + s_\theta(X_{t-1}, \ldots X_{t-q}) \cdot e_t, \, t \in \mathbb{Z}. \tag{6}$$

In case, the data-generating process does not belong to class (5) or (6), a residual bootstrap making use of such a model fit asymptotically can only work if the asymptotic distribution of the parameter of interest does not vary if switching from the true underlying process to a process of type (5) or (6), respectively.

The simplest situation in this context one might think of is a causal (linear) autoregressive model of fixed and known order p and with i.i.d. innovations (e_t) (having zero

mean and at least finite second-order moments) for the data-generating process, i.e.,

$$X_t = \sum_{k=1}^{p} a_k X_{t-k} + e_{t-k}, t \in \mathbb{Z}. \tag{7}$$

Of course in such a situation, it suffices to consider an autoregressive process of the same order p with consistently estimated parameters $\widehat{a}_k$ (e.g., Yule-Walker estimates) and consistently estimated distribution of the innovations in the bootstrap world. If the statistic of interest is the centered autocovariance or centered autocorrelation function evaluated at some lags, then it is known that the asymptotic distribution for these quantities is not the same for linear AR(p) processes of type (7) and, for example, general mixing processes. This means that the residual bootstrap based on an autoregressive fit in general does not lead to consistent results.

As long as one is interested in the distribution of the coefficients of the (linear) autoregressive fit itself and as long as the underlying model follows (7), even the wild bootstrap proposal (4) leads to valid approximation results. The bootstrap estimators in such a situation just are the coefficients of a linear regression of X_t^* on $X_{t-1}, \ldots, X_{t-p}$. The reason is that the asymptotic distributions of Yule-Walker and least-squares estimators for the coefficients in linear autoregression and linear regression with i.i.d. errors coincide. For more general statistics, it is of course not true that the wild bootstrap proposal (4) leads to asymptotically valid results, because in the bootstrap world, we even do not generate a stochastic process.

The application of a residual resampling scheme (2) in principle is of course not limited to causal (linear) autoregressive processes but easily can be extended to a broad class of further parametric models (including ARMA, threshold, ARCH, and GARCH models). Relevant references for ARMA models are Bose (1988), Bose (1990), and Franke and Kreiss (1992). The multivariate ARMA situation is considered in Paparoditis and Streitberg (1991). Basawa et al. (1991), Datta (1996), and Heimann and Kreiss (1996) dealt with the situation of general AR(1) models in which the parameter value is not restricted to the stationary case. For first-order autoregressions with positive innovations, Datta and McCormick (1995a) considered a bootstrap proposal for an estimator specific to the considered situation. Finally, Franke et al. (2006) considered the application of the bootstrap to order selection in autoregression, and Paparoditis and Politis (2005) considered bootstrap methods for unit root testing in autoregressions. It is worth mentioning that the assumption of i.i.d. innovations is rather essential for the asymptotic validity of the described bootstrap proposals for most statistics of interest. For a bootstrap test for a unit root in autoregressions with weakly dependent errors, see Psaradakis (2001).

Finally, let us come back to the fully nonparametric situation. If the data-generating process follows a nonparametric model equation of the form

$$X_t = m(X_{t-1}, \ldots, X_{t-p}) + s(X_{t-1}, \ldots, X_{t-q}) \cdot e_t, t \in \mathbb{Z}, \tag{8}$$

again with i.i.d. innovations (e_t) (having zero mean and unit variance) and known orders p, q, in order to define a bootstrap process according to (2) or (4), we have to apply nonparametric estimators of the underlying functional parameters $m : \mathbb{R}^p \to \mathbb{R}$ and $s : \mathbb{R}^q \to [0, \infty]$, which are conditional mean and conditional volatility function of

the process. For smooth mean functions m and smooth volatility functions v, kernel-based estimators successfully could be applied, while for more general situations, wavelet-based estimators may be used. It can be expected that for almost all statistical quantities, a residual bootstrap based on a nonparametric model fit for (8) will lead to a consistent resampling procedure.

As far as nonparametric estimators are of interest, one can take advantage of the so-called *whitening by windowing* effect, which in many situations of interest implies that the dependence structure of the underlying process does not show up in asymptotic distributions of nonparametric estimates. Because of this, one might also take regression-type standard as well as wild residual bootstrap procedures like (4) into consideration, which are often much easier to implement because they completely ignore the underlying dependence structure. We refer to Franke et al. (2002a) and Franke et al. (2002b) for nonparametric kernel-based-bootstrap methods. Neumann and Kreiss (1998) and Kreiss (2000) considered to what extent the nonparametric regression type bootstrap procedures successfully can be applied to situations (8) as long as nonparametric estimators and tests for conditional mean and/or volatility functions in nonparametric autoregressions are considered. A local bootstrap approach to kernel estimation for dependent observations is suggested and investigated in Paparoditis and Politis (2000).

Nonparametric bootstrap applications to goodness-of-fit testing problems for mean and volatilty functions in models of the form (8) are derived and discussed in Kreiss and Neumann (1999) and Kreiss et al. (2008). Paparoditis and Politis (2003) applied the concept of block bootstrap (cf. Section 5) to residuals in order to deal with rather relevant unit root testing problems.

3. Autoregressive-sieve bootstrap

The main idea of autoregressive (AR)-sieve bootstrap follows the lines of residual bootstrap described in Section 2. Instead of applying the *drawing with replacement* idea to residuals of an in some sense optimal predictor, we restrict for the AR-sieve bootstrap to (optimal) linear predictors, given an increasing number of past values of the underlying process itself.

If we again assume that the underlying process is stationary and, moreover, has positive variance $\gamma(0) > 0$ and asymptotically (as $h \to \infty$) vanishing autocovariances $\gamma(h)$, then we obtain from Brockwell and Davis (1991), Prop. 5.1.1, that the matrix $\mathbf{\Gamma}_p = (\gamma(i-j))_{i,j=1,2,\ldots,p}$ is positive definite, and therefore, immediately the best (in mean square sense) linear predictor of X_{j+1} given p past values $\mathbf{X}_{j,p} = (X_j, \ldots, X_{j-p+1})$ exists, which is unique and is given by $\widehat{X}_{j+1} = \sum_{j=1}^{p} a_j(p) X_{t-j}$. The coefficients $(a_j(p) j = 1, 2, \ldots, p)$ efficiently can be calculated from

$$(a_1(p), a_2(p), \ldots, a_p(p))^T = \mathbf{\Gamma}_p^{-1}(\gamma(1), \gamma(2), \ldots, \gamma(p))^T.$$

Now, one way to generate bootstrap pseudo-time series is to select a set of p starting values $X_1^*, X_2^*, \ldots, X_p^*$ and, given the past $X_1^*, X_2^*, \ldots, X_j^*$, $j \geq p$, to generate the next observation X_{j+1}^* using an estimated version of the best linear predictor $\widehat{X}_{j+1} = \sum_{s=1}^{p} a_s(p) X_{j+1-s}^*$ plus an error term which is selected randomly from the set of centered estimated prediction errors $X_{t+1} - \widehat{X}_{t+1} = X_{t+1} - \sum_{s=1}^{p} a_s(p) X_{t+1-s}$. This idea together with the order p converging to infinity as sample size n increases

lead to the so-called AR-sieve bootstrap procedure, which can be summarized in the following steps.

Step 1: Select an order $p = p(n) \in \mathbb{N}$, $p \ll n$, and fit a pth order autoregressive model to $X_1, X_2, \ldots, X_n$. Denote by $\widehat{a}(p) = (\widehat{a}_j(p), j = 1, 2, \ldots, p)$, the Yule-Walker autoregressive parameter estimators, that is $\widehat{a}(p) = \widehat{\Gamma}(p)^{-1}\widehat{\gamma}_p$, where for $0 \le h \le p$,

$$\widehat{\gamma}_X(h) = \frac{1}{n}\sum_{t=1}^{n-|h|}(X_t - \overline{X}_n)(X_{t+|h|} - \overline{X}_n),$$

$\overline{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$, $\widehat{\Gamma}(p) = (\widehat{\gamma}_X(r-s))_{r,s=1,2,\ldots,p}$ and $\widehat{\gamma}_p = (\widehat{\gamma}_X(1), \ldots, \widehat{\gamma}_X(p))'$.

Step 2: Let $\widetilde{\varepsilon}_t(p) = X_t - \sum_{j=1}^{p}\widehat{a}_j(p)X_{t-j}$ $t = p+1, p+2, \ldots, n$, be the residuals of the autoregressive fit and denote by $\widehat{F}_n$ the empirical distribution function of the centered residuals $\widehat{\varepsilon}_t(p) = \widetilde{\varepsilon}_t(p) - \overline{\varepsilon}$, where $\overline{\varepsilon} = (n-p)^{-1}\sum_{t=p+1}^{n}\widetilde{\varepsilon}_t(p)$

Let $(X_1^*, X_2^*, \ldots, X_n^*)$ be a set of observations from the time series $\mathbf{X}^* = \{X_t^* : t \in \mathbb{Z}\}$, where $X_t^* = \sum_{t=1}^{p}\widehat{a}_j(p)X_{t-j}^* + e_t^*$ and the e_t^*'s are independent random variables having identical distribution $\widehat{F}_n$.

Step 3: Let $T_n^* = T_n(X_1^*, X_2^*, \ldots, X_n^*)$ be the same estimator as the estimator T_n of interest based on the pseudo-time series $X_1^*, X_2^*, \ldots, X_n^*$, and ϑ^* the analogue of ϑ associated with the bootstrap process $\mathbf{X}^*$. The AR-sieve bootstrap approximation of $\mathcal{L}_n = \mathcal{L}(c_n(\hat{\theta}_n - \theta))$ is then given by $\mathcal{L}_n^* = \mathcal{L}^*(c_n(T_n^* - \vartheta^*))$.

Using Yule-Walker estimators in Step 1 of the AR-sieve bootstrap is rather convenient. Besides simple, stable, and fast computation (using the Durbin–Levinson algorithm), it ensures that the complex polynomial $\widehat{A}_p(z) = 1 - \sum_{j=1}^{p}\widehat{a}_j(p)z^j$ has no roots on or within the unit disc $\{z \in \mathbb{C} : |z| \le 1\}$, i.e., the bootstrap process $\mathbf{X}^*$ is always a stationary and causal autoregressive process (cf. Kreiss and Neuhaus (2006), Satz 8.7 and Bemerkung 8.8).

The described AR-sieve bootstrap has been introduced by Kreiss (1988) and has been investigated from several points of view in Paparoditis and Streitberg (1991), Kreiss (1992), Paparoditis (1996), Bühlmann (1997), Kreiss (1997), Bühlmann (1998), Choi and Hall (2000), Gonçalves and Kilian (2007), Poskitt (2008), and recently in Kreiss et al. (2011). Park (2002) gives an invariance principle for the sieve bootstrap and Bose (1988) worked out the edgeworth correction of bootstrap in autoregressions. Kapetanios (2010) applied the idea of sieve bootstrap to long-memory processes.

The question of course is under what assumptions on the underlying stochastic process $(X_t : t \in \mathbb{Z})$ and for what kind of statistics $T_n(X_1, \ldots, X_n)$ can we successfully approximate the distribution $\mathcal{L}_n$ by that of $\mathcal{L}_n^*$? In almost all papers concerning AR-sieve bootstrap, it is assumed that (X_t) is a *linear* autoregression of possibly infinite order, i.e.,

$$X_t = \sum_{j=1}^{\infty} a_j X_{t-j} + e_t, \tag{9}$$

with (e_t) an i.i.d. sequence and absolutely summable coefficients a_j, which moreover typically are assumed to decrease polynomially or even exponentially fast. An exception is the sample mean $\overline{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$, where Bühlmann (1997) showed that for this

specific statistic, the assumption of i.i.d. innovations (e_t) can be relaxed to martingale differences.

Kreiss et al. (2011) used the fact that every purely nondeterministic, zero mean stationary process possessing a strictly positive and continuous spectral density has a unique Wold-type autoregressive representation of the form

$$X_t = \sum_{j=1}^{\infty} a_j X_{t-j} + \varepsilon_t, \tag{10}$$

with absolutely summable coefficients a_k and a white noise process (ε_t) consisting of zero mean, uncorrelated random variables. The representation (10) does by far not mean that the underlying process is a linear, causal AR(∞) process driven by i.i.d. innovations!

Kreiss et al. (2011) have shown that under rather mild regularity assumptions, the AR-sieve bootstrap asymptotically correctly mimics the behavior of the following so-called companion autoregressive process $(\widetilde{X}_t : t \in \mathbb{Z})$ defined according to

$$\widetilde{X}_t = \sum_{j=1}^{\infty} a_j \widetilde{X}_{t-j} + \widetilde{\varepsilon}_t, \tag{11}$$

where the innovation process $(\widetilde{\varepsilon}_t)$ consists of i.i.d. random variables whose marginal distribution coincides with that of (ε_t), i.e., $\mathcal{L}(\varepsilon_t) = \mathcal{L}(\widetilde{\varepsilon}_t)$ and the coefficients are those of the Wold-type autoregressive representation (10). Note that the first- and second-order properties of the two stochastic processes $(\widetilde{X}_t)$ and (X_t) are the same, i.e., autocovariances and the spectral density coincide. However, all probability characteristics beyond second-order quantities are not necessarily the same and, in general, will substantially differ. Kreiss et al. (2011) showed for a rather general class of statistics that the AR-sieve bootstrap asymptotically works if the asymptotic distribution of the statistics of interest is the same for the underlying process (X_t) and the companion autoregressive process $(\widetilde{X}_t)$. This rather plausible check criterion for the AR-sieve bootstrap to work leads, for example, for the arithmetic mean under very mild assumptions (much weaker than martingale differences for the innovations) to consistency of the AR-sieve proposal. For autocorrelations, this check criterion shows that AR-sieve bootstrap works if the underlying process possesses any linear representation with i.i.d. errors not depending on whether this representation can be inverted to an AR(∞)–representation with i.i.d. errors or not. For further details, we refer to Kreiss et al. (2011).

4. Bootstrap for Markov chains

Extension of the Bootstrap methods from i.i.d. random variables to Markov chains was initiated by Kulperger and Prakasa Rao (1989) for the finite state space case. Suppose that $\{X_n\}_{n\geq 0}$ be a stationary Markov chain with a finite state space $S = \{s_1, \ldots, s_\ell\}$, where $\ell \in \mathbb{N}$ and where $\mathbb{N} \equiv \{1, 2, \ldots\}$ denotes the set of all natural integers. Let the $\ell \times \ell$ transition probability matrix of the chain be given by $\mathbb{P} = ((p_{ij}))$ and the stationary distribution by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_\ell)$. Thus, for any $1 \leq i, j \leq \ell$, $p_{ij} = P(X_1 = s_j | X) = s_i)$

and $\pi_i = P(X_0 = s_i)$. The joint distribution of the chain is completely determined by the finitely many unknown parameters, given by the components of $\boldsymbol{\pi}$ and $\mathbb{P}$. Given a sample $X_0, \ldots, X_{(n-1)}$ of size n from the Markov chain, we can estimate the population parameters π_i's and p_{ij}'s as

$$\hat{\pi}_i = n^{-1} \sum_{k=0}^{n-1} \mathbb{1}(X_k = s_i) \quad \hat{p}_{ij} = n^{-1} \sum_{k=0}^{n-2} \mathbb{1}(X_k = s_i, X_{k+1} = s_j)/\hat{\pi}_i, \quad (12)$$

$1 \le i, j \le \ell$. The bootstrap observations $X_0^*, \ldots, X_{n-1}^*$ can now be generated using the estimated transition matrix and the marginal distribution. Specifically, first generate a random variable X_0^* from the discrete distribution on $\{1, \ldots, \ell\}$ that assigns mass $\hat{\pi}_i$ to s_i, $1 \le i \le \ell$. Next, having generated $X_0^*, \ldots, X_{k-1}^*$ for some $1 \le k < n-1$, generate X_k^* from the discrete distribution on $\{1, \ldots, \ell\}$ that assigns mass $\hat{p}_{ij}$ to j, $1 \le j \le \ell$, where s_i is the value of X_{k-1}^*. The bootstrap version of a given random variable $T_n = t_n(\mathbf{X}_n; \theta)$ based on $(X_0, \ldots, X_{n-1})$ and a parameter θ of interest is now defined as

$$T_n^* = t_n(X_0^*, \ldots, X_{n-1}^*; \hat{\theta}_n)$$

where $\hat{\theta}_n$ is an estimator of θ based on $X_0, \ldots, X_{n-1}$. For example, for $T_n = n^{1/2}(\bar{X}_n - \mu)$, where $\bar{X}_n = n^{-1} \sum_{k=0}^{n-1} X_i$ and $\mu = EX_0$, we set $T_n^* = n^{1/2}(\bar{X}_n^* - \hat{\mu}_n)$, where $\bar{X}_n^*$ is the average of the n bootstrap variables X_k^*'s and where $\hat{\mu}_n = \sum_{i=1}^{\ell} \hat{\pi}_i X_i$, the (conditional) expectation of X_0^* given $\mathbf{X}_n$. This approach has been extended to the countable case by Athreya and Fuh (1992).

More recently, different versions of the Bootstrap method for Markov processes based on estimated transition probability functions have been extended to the case, where the state space is Euclidean. In this case, one can use the nonparametric function estimation methodology to estimate the marginal distribution and the transition probability function. For consistency of the method, see Rajarshi (1990), and for the second-order properties of the method, see Horowitz (2003). A "local" version of the method (called the Local Markov Bootstrap or MLB, in short) has been put forward by Paparoditis and Politis (2001b). The idea here is to construct the bootstrap chain by sequential drawing – having selected a set of bootstrap observations, the next observation is randomly selected from a "neighborhood of close values" of the observation(s) in the immediate past. Paparoditis and Politis (2001b) showed that the resulting bootstrap chain was stationary and Markov and also that it enjoyed some robustness with regard to the Markovian assumption. For more on the properties of the MLB, see Paparoditis and Politis (2001b).

A completely different approach to bootstrapping Markov chains was introduced by Athreya and Fuh (1992). Instead of using estimated transition probabilities, they formulate a resampling scheme based on the idea of regeneration. A well-known result (Athreya and Ney, 1978) on Markov chains literature says that for a large class of Markov chains satisfying the so-called *Harris recurrence condition*, successive returns to a recurrent state gives a decomposition of the chain into i.i.d. cycles (of random lengths). The regeneration-based bootstrap resamples these i.i.d. cycles to generate the bootstrap observations. Here, we describe it for a Markov Chain $\{X_n\}_{n \ge 0}$ with values in a general state space S, equipped with a countably generated σ-field $\mathcal{S}$. Let $\mathbb{P}(x, dy)$

denote the transition probability function, and let $\pi(\cdot)$ denote the stationary distribution of the Markov chain. Suppose that $\{X_n\}_{n\geq 0}$ is positive recurrent with a known "accessible atom" $A \in \mathcal{S}$; Here, a set $A \in \mathcal{S}$ is called an "accessible atom" if it satisfies

$$\pi(A) > 0 \quad \text{and} \quad \mathbb{P}(x,\cdot) = \mathbb{P}(y,\cdot) \quad \text{for all} \quad x, y \in A.$$

For a Harris recurrent Markov chain with a countable state space, this condition holds trivially. Define the successive return times to A by

$$\tau_1 = \inf\{m \geq 1 : X_m \in A\} \quad \text{and}$$
$$\tau_{k+1} = \inf\{m \geq \tau_k : X_m \in A\},\ k \geq 1.$$

Then, by strong Markov property, the blocks $\mathbb{B}_k = \{X_i : \tau_k + 1 \leq i \leq \tau_{k+1}\}, k \geq 1$ are i.i.d. variables with values in the taurus $\cup_{k\geq 1} S^k$. The *regeneration-based bootstrap* resamples the collection of blocks

$$\left\{\mathbb{B}_k : \mathbb{B}_k \subset \{X_0, \ldots, X_{n-1}\}\right\}$$

with replacement to generate the bootstrap observations. Validity of the method for the sample mean in the countable state space case is established by Athreya and Fuh (1992). For second-order properties of the regeneration-based bootstrap, see Datta and McCormick (1995b), and its refinements in Bertail and Clemencon (2006). Bertail and Clemencon (2006) show that the regeneration-based bootstrap, with a proper definition of the bootstrap version, achieves almost the same level of accuracy as in the case of i.i.d. random variables for linear statistics. As a result, for Markov chains satisfying the requisite regularity conditions, one should use the regeneration-based bootstrap (with blocks of random lengths) instead of the block bootstrap methods described below which are applicable to more general processes but are not as accurate.

5. Block bootstrap methods

For time series that are *not* assumed to have a specific structural form, Künsch (1989) formulated a general bootstrap method, currently known as the *moving block bootstrap* or MBB, in short. Quite early in the bootstrap literature, Singh (1981) showed that resampling single observations, as considered by Efron (1979) for independent data, failed to produce valid approximations in presence of dependence. As a remedy for the limitation of the single-data-value resampling scheme for dependent time series data, Künsch (1989) advocated the idea of resampling blocks of observations at a time (see also Bühlmann and Künsch (1995)). By retaining the neighboring observations together within the blocks, the dependence structure of the random variables at short lag distances is preserved. As a result, resampling blocks allows one to carry this information over to the bootstrap variables. The same resampling plan was also independently suggested by Liu and Singh (1992), who coined the term "moving block bootstrap."

We now briefly describe the MBB. Suppose that $\{X_t\}_{t\in\mathbb{N}}$ is a stationary weakly dependent time series and that $\{X_1, \ldots, X_n\} \equiv \mathbf{X}_n$ are observed. Let ℓ be an integer

satisfying $1 \le \ell < n$. Define the overlapping blocks $\mathbb{B}_1, \ldots, \mathbb{B}_N$ of length ℓ contained in $\mathbf{X}_n$ as

$$\begin{aligned}
\mathbb{B}_1 &= (X_1, X_2, \ldots, X_\ell), \\
\mathbb{B}_2 &= \quad (X_2, \ldots, X_\ell, X_{\ell+1}), \\
&\ldots \qquad\qquad\qquad\qquad\qquad \ldots \\
\mathbb{B}_N &= \qquad\qquad\qquad\qquad\qquad (X_{n-\ell+1}, \ldots, X_n),
\end{aligned}$$

where $N = n - \ell + 1$. For simplicity, suppose that ℓ divides n. Let $b = n/\ell$. To generate the MBB samples, we select b blocks at random with replacement from the collection $\{\mathbb{B}_1, \ldots, \mathbb{B}_N\}$. Since each resampled block has ℓ elements, concatenating the elements of the b resampled blocks serially yields $b \cdot \ell$ bootstrap observations $X_1^*, \ldots, X_n^*$. Note that if we set $\ell = 1$, then the MBB reduces to the ordinary bootstrap method of Efron (1979) for i.i.d. data. However, for a valid approximation in the dependent case, it is typically required that

$$\ell^{-1} + n^{-1}\ell = o(1) \quad \text{as} \quad n \to \infty. \tag{13}$$

Some typical choices of ℓ are $\ell = Cn^{1/k}$ for $k = 3, 4$, where $C \in \mathbb{R}$ is a constant. Next, suppose that the random variable of interest is of the form $T_n = t_n(\mathbf{X}_n; \theta(P_n))$, where $P_n = \mathcal{L}(\mathbf{X}_n)$ denotes the joint probability distribution of $\mathbf{X}_n$. The MBB version of T_n based on blocks of size ℓ is defined as

$$T_n^* = t_n(X_1^*, \ldots, X_n^*; \theta(\hat{P}_n)),$$

where $\hat{P}_n = \mathcal{L}(X_1^*, \ldots, X_n^* | \mathbf{X}_n)$, the conditional joint probability distribution of $X_1^*, \ldots, X_n^*$, given $\mathbf{X}_n$, and where we suppress the dependence on ℓ to ease the notation. In the general case, where n is not a multiple of ℓ, one may resample $b = b_0$ blocks, where $b_0 = \min\{k \ge 1 : k\ell \ge n\}$ and retain the first n resampled data-values to define the bootstrap replicate of T_n.

To illustrate the construction of T_n^* in a specific example, suppose that T_n is the centered and scaled sample mean $T_n^{1/2}(\bar{X}_n - \mu)$. Then, the MBB version of T_n is given by $T_n^* = n^{1/2}(\bar{X}_n^* - \tilde{\mu}_n)$, where $\bar{X}_n^*$ is the sample mean of the bootstrap observations and where $\tilde{\mu}_n = E_*(\bar{X}_n^*)$. It is easy to check that

$$\begin{aligned}
\tilde{\mu}_n &= N^{-1} \sum_{i=1}^{N} \big(X_i + \cdots + X_{i+\ell-1}\big)/\ell \\
&= N^{-1} \left[\sum_{i=\ell}^{N} X_i + \sum_{i=1}^{\ell-1} \frac{i}{\ell} \big(X_i + X_{n-i+1}\big) \right],
\end{aligned} \tag{14}$$

which is different from $\bar{X}_n$ for $\ell > 1$. Lahiri (1991) established second-order correctness of the MBB approximation for the normalized sample mean, where the bootstrap sample mean is centered at $\tilde{\mu}_n$. The "naive" centering of $\bar{X}_n^*$ at $\bar{X}_n$ is not appropriate as it leads to a loss of accuracy of the MBB approximation (Lahiri, 1992). Second-order

correctness of the MBB approximation for studentized statistics has been established independently by Götze and Künsch (1996) for stationary processes and by Lahiri (1996) in multiple linear regression models with dependent errors.

Several variants of the block bootstrap method exist in the literature. One of the early versions of the block bootstrap, implicit in the work of Carlstein (1986), restricts attention to the collection of nonoverlapping blocks in the data, and resamples from this smaller collection to generate the bootstrap observations. This is known as the *nonoverlapping block bootstrap* (NBB). To describe it briefly, suppose that ℓ is an integer in $(1, n)$ satisfying (13). Also, for simplicity, suppose that ℓ divides n and set $b = n/\ell$. The NBB samples are generated by selecting b blocks at random with replacement from the collection $\{\tilde{\mathbb{B}}_1, \ldots, \tilde{\mathbb{B}}_b\}$, where

$$\begin{aligned}
\tilde{\mathbb{B}}_1 &= (X_1, \ldots, X_\ell), \\
\tilde{\mathbb{B}}_2 &= \qquad\qquad (X_{\ell+1}, \ldots, X_{2\ell}), \\
&\cdots \qquad\qquad\qquad\qquad\qquad\qquad \cdots \\
\tilde{\mathbb{B}}_b &= \qquad\qquad\qquad\qquad\qquad\qquad (X_{(b-1)\ell+1}, \ldots, X_n).
\end{aligned}$$

Because the blocks in the NBB construction do not overlap, it is easier to analyze theoretical properties of NBB estimators than those of MBB estimators of a population parameter. However, the NBB estimators typically have higher MSEs at any block size ℓ compared to their MBB counterparts (cf. Lahiri (1999)).

Other variants of the block bootstrap include the *circular block bootstrap* (CBB) and the *stationary bootstrap* (SB) of Politis and Romano (1992, 1994), the *matched block bootstrap* (MaBB) of Carlstein et al. (1998), the *tapered block bootstrap* (TBB) of Paparoditis and Politis (2001a), among others. The CBB and the SB are primarily motivated by the need to remove the uneven weighting of the observations at the beginning and at the end in the MBB (cf. (14)) and are based on the idea of periodic extension of the observed segment of the time series. Further, while most block bootstrap methods are based on blocks of a deterministic length ℓ, the SB is based on blocks of random lengths that have a Geometric distribution with expected length ℓ satisfying (13). The biases of the variance estimators generated by the MBB, NBB, CBB, and SB are of the order $O(\ell^{-1})$, while the variances are of the order $O(n^{-1}\ell)$, where ℓ denotes the block size and n the sample size. It turns out that the MBB and the CBB have asymptotically equivalent performance and are also the most accurate of these four methods. For relative merits of these four methods, see Lahiri (1999), Politis and White (2004), and Nordman (2009). The MaBB uses a stochastic mechanism to reduce the edge effects from joining independent blocks in the MBB, while the TBB shrinks the boundary values in a block towards a common value, like the sample mean, to achieve the same. Although somewhat more complex than the MBB or the CBB, both the MaBB and the TBB yield more accurate variance estimators, with biases of the order $O(\ell^{-2})$ and variances of the order $O(n^{-1}\ell)$. In this sense, both MaBB and TBB are considered second-generation block bootstrap methods.

Performance of the block bootstrap methods crucially depends on the choice of the block size and on the dependent structure of the process. Explicit formulas for MSE-optimal block sizes for estimating the variances of smooth functions of sample means are known for the MBB, CBB, NBB, and SB (Hall et al., 1995; Lahiri, 1999). Thus,

one can use these expressions to formulate plug-in estimators of the optimal block sizes (Patton et al., 2009; Politis and White, 2004). For the variance estimation problem, Bühlmann and Künsch (1999) formulated a method based on linearization of an estimator using its influence function, which is somewhat more general than the direct plug-in approach. But perhaps the most widely used method in this context is given by Hall et al. (1995) who develop a general empirical method for estimating the optimal block sizes for estimating *both* the variance and the distribution function. The Hall et al. (1995) method uses the subsampling method to construct an estimator of the MSE as a function of the block size, and then minimize it to produce the estimator of the optimal block size. An alternative method based on the Jackknife-after-bootstrap method (Efron, 1992; Lahiri, 2002) has been recently proposed by Lahiri et al. (2007). They call it a *nonparametric plug-in* (NPPI) method, as it works like a plug-in method, but at the same time, it does not require the user to find an exact expression for the optimal block size analytically. The key construction of the NPPI method combines more than one resampling method suitably and, thereby, implicitly estimates the population parameters that appear in the formulas for the optimal block sizes. Further, the NPPI method is applicable to block bootstrap estimation problems involving the variance, the distribution function, and the quantiles. However, it is a computationally intensive method as it uses a combination of bootstrap and Jackknife methods.

For further discussion of the block length selection rules for block bootstrap methods, see Lahiri (2003a, Chapter 7) and the references therein.

6. Frequency domain bootstrap methods

An alternative bootstrap method that completely avoids the difficult problem of block length selection is given by the *Frequency Domain Bootstrap* (FDB).

One can apply the FDB for inference on population parameters of a second-order stationary process that can be expressed as a functional of its spectral density. Here, we give a short description of the FDB (see Paparoditis (2002) for an overview on frequency domain bootstrap methods). Given the data $\mathbf{X}_n$, define its Fourier transform

$$Y_n(w) = n^{-1/2} \sum_{t=1}^{n} X_t \exp(-\iota w t), \quad w \in (-\pi, \pi]. \tag{15}$$

The formulation of the FDB is based on the following well-known results:

(i) the Fourier transforms $Y_n(\lambda_1), \ldots, Y_n(\lambda_k)$ are *asymptotically independent* for any set of distinct ordinates $-\pi < \lambda_1 < \cdots < \lambda_k \le \pi$ (cf. Brockwell and Davis (1991), Lahiri (2003b));

(ii) The original observations $\mathbf{X}_n$ admit a representation in terms of the transformed values $\mathbf{Y}_n = \{Y_n(w_j) : j \in \mathcal{I}_n\}$ as (cf. Brockwell and Davis (1991)),

$$X_t = n^{-1/2} \sum_{j \in \mathcal{I}_n} Y_n(w_j) \exp(\iota t w_j), \quad t = 1, \ldots, n \tag{16}$$

where $\iota = \sqrt{-1}$, $w_j = 2\pi j/n$, and $\mathcal{I}_n = \{-\lfloor (n-1) \rfloor/2, \ldots, \lfloor (n-1) \rfloor/2\}$.

Thus, one can express a given variable $R_n = r_n(\mathbf{X}_n; \theta)$ also in terms of the transformed values $\mathbf{Y}_n$ and resample from the Y-values to define the FDB version of R_n. Variants of the FDB method have been proposed and studied by Hurvich and Zeger (1987) and Franke and Härdle (1992). Under some regularity conditions, Dahlhaus and Janas (1996) established second-order correctness of the FDB for a class of estimators called the "ratio statistics." Ratio statistics are defined as the ratio of two "spectral mean" estimators of the form $\int_0^\pi g(w) I_n(w) dw$, where $g : [0, \pi) \to \mathbb{R}$ is an integrable function and where $I_n(w) = |Y(w)|^2$ is the periodogram of $\mathbf{X}_n$. A common example of a ratio estimator is the lag-k sample autocorrelation coefficient, $k \geq 1$, given by

$$\hat{\rho}_n(k) = r_n(k)/r_n(0),$$

where, for any $m \geq 0$, $r_n(m) = n^{-1} \sum_{i=1}^{n-m} X_i X_{i+m}$ is a (mean-uncorrected) version of the sample autocovariance function at lag m. It is easy to check that $r_n(m) = 2 \int_0^\pi \cos(mw) I_n(w) dw$, and therefore, $\hat{\rho}_n(k)$ is a ratio-statistic estimating the population kth order lag autocorrelation coefficient $\rho(k) = EX_1 X_{1+k}/EX_1^2$, when $\{X_n\}$ is a zero-mean second-order stationary process.

Although the FDB avoids the problem of block length selection, second-order accuracy of the FDB distributional approximations is available only under restrictive regularity conditions (cf. Dahlhaus and Janas (1996)). Further, it is known (cf. Lahiri (2003a, Section 9.2)) that accuracy of the FDB for spectral means and ratio estimators is rather sensitive to deviations from the model assumptions. Frequency domain bootstrap methods can also be applied to testing problems, cf. Dette and Paparoditis (2009).

Paparoditis and Politis (1999) applied the idea of a localized bootstrap approach to periodogram statistics, while a more general version of the FDB is proposed by Kreiss and Paparoditis (2003), which adds an intermediate autoregressive model fitting step in an attempt to capture higher order cross-cumulants of the DFTs. Kreiss and Paparoditis (2003) show that the modified version of the FDB provides a valid approximation for a wider class of spectral mean estimators that includes the class of ratio estimators covered by the FDB. We elaborate on this in the next section.

7. Mixture of two bootstrap methods

So far, we discussed several bootstrap proposals which are either defined in time domain (like block-, residual, AR-sieve and Markovian bootstrap) or defined in frequency domain (like periodogram-bootstrap). In this section, we briefly discuss mixtures of two bootstrap proposals (so-called hybrid bootstrap procedures). The rational behind such proposals is to bring together advantages of resampling approaches from both fields.

The hybrid bootstrap procedure proposed in Kreiss and Paparoditis (2003) can be understood as an extension of AR-sieve bootstrap as well as an extension of frequency domain bootstrap. As described in Section 3, AR-sieve bootstrap uses an autoregressive fit in order to obtain residuals of this fit. It can be argued that these residuals under reasonably assumptions on the data-generating process can be regarded to behave

approximately like i.i.d. random variables. Since such an i.i.d. property for the residuals does (if at all) at most holds approximately, it might be advisable to add a further nonparametric step to the AR-sieve bootstrap which is able to correct for data features which cannot or are not represented by the autoregressive fit.

On the other hand, frequency domain bootstrap as described above mainly uses the fact that periodogram ordinates asymptotically behave like i.i.d. random variables. But neglecting the existing and only asymptotically vanishing dependence structure between contiguous periodogram ordinates leads to drawbacks of frequency domain bootstrap. Therefore, an additional step of fitting a parametric model (e.g., an autoregressive model) to the data and applying – in the spirit of Tukey's pre-whitening – a frequency domain bootstrap approach to the residuals of the fit partly is able to remove this remedy. If, for example, the true underlying spectral density has some dominant peaks, then pre-whitening leads to a considerable improvement of nonparametric spectral density estimators. An autoregressive fit really is able to catch the peaks of the spectral density rather well and the curve $I_n(\lambda)/\hat{f}_{\mathrm{AR}}(\lambda)$, cf. Step 5 below, is much smoother than $I_n(\lambda)$, thus much easier to estimate nonparametrically.

Based on this motivation, an autoregressive-aided frequency domain hybrid bootstrap can be described along the following five steps. It is worth mentioning that fitting an autoregression should be understood as a (convenient) example. Of course, fitting other parametric models may be regarded as a pre-stage of frequency domain bootstrap.

Step 1: Given the observations $X_1, \ldots, X_n$, we fit an autoregressive process of order p, where p may depend on the particular sample at hand.

This leads to estimated parameters $\hat{a}_1(p), \ldots, \hat{a}_p(p)$ and $\hat{\sigma}(p)$, which are obtained from the common Yule-Walker equations. Consider the estimated residuals

$$\hat{\varepsilon}_t = X_t - \sum_{\nu=1}^{p} \hat{a}_\nu(p) X_{t-\nu},\ t = p+1, \ldots, n,$$

and denote by $\hat{F}_n$ the empirical distribution of the standardized quantities $\widehat{\varepsilon}_{p+1}, \ldots, \widehat{\varepsilon}_n$, i.e., $\widehat{F}_n$ has mean zero and unit variance.

Step 2: Generate bootstrap observations $X_1^+, X_2^+, \ldots, X_n^+$, according to the following autoregressive model of order p

$$X_t^+ = \sum_{\nu=1}^{p} \widehat{a}_\nu(p) X_{t-\nu}^+ + \widehat{\sigma}(p) \cdot \varepsilon_t^+ ,$$

where (ε_t^+) constitutes a sequence of i.i.d. random variables with cumulative distribution function $\widehat{F}_n$ (conditionally on the given observations $X_1, \ldots, X_n$).

The bootstrap process $\mathbf{X}^+ = (X_t^+ : t \in \mathbb{Z})$ possesses the following spectral density:

$$\hat{f}_{\mathrm{AR}}(\lambda) = \frac{\hat{\sigma}^2(p)}{2\pi} \left| 1 - \sum_{\nu=1}^{p} \hat{a}_\nu(p) e^{-i\nu\lambda} \right|^{-2} ,\ \lambda \in [0, \pi].$$

Note that because we make use of the Yule-Walker parameter estimators in Step 1, it is always ensured that $\hat{f}_{\mathrm{AR}}$ is well-defined, i.e., the polynomial

$1 - \sum_{v=1}^{p} \hat{a}_v(p) z^v$ has no complex roots with magnitude less than or equal to one. Moreover, the bootstrap autocovariances $\gamma^+(h) = E^+ X_1^+ X_{1+h}^+, h = 0, 1, \ldots, p$ coincide with the empirical autocovariances $\hat{\gamma}_n(h)$ of the underlying observations. It should be noted that it is convenient, but not necessary to work with Yule-Walker parameter estimates. Any $\sqrt{n}$-consistent parameter estimates would suffice.

Step 3: Compute the periodogram of the bootstrap observations, i.e.,

$$I_n^+(\lambda) = \frac{1}{2\pi n} \left| \sum_{t=1}^{n} X_t^+ e^{-i\lambda t} \right|^2 , \lambda \in [0, \pi].$$

Step 4: Define the following nonparametric estimator $\widehat{q}$

$$\widehat{q}(\lambda) = \frac{1}{n} \sum_{j=-N}^{N} K_h\Big(\lambda - \lambda_j\Big) \frac{I_n(\lambda_j)}{\widehat{f}_{\mathrm{AR}}(\lambda_j)} , \quad \text{for} \quad \lambda \in [0, \pi),$$

while for $\lambda = \pi$, $\hat{q}(\pi)$ is defined as twice the quantity on the right-hand side of the above equation taking into account that no Fourier frequencies greater than π exist. Here and above, the λ_j's denote the Fourier frequencies, $K : [-\pi, \pi] \to [0, \infty)$ denotes a probability density (kernel), $K_h(\cdot) = h^{-1} K(\cdot / h)$, and $h > 0$ is the so-called bandwidth.

Step 5: Finally, the bootstrap periodogram I_n^* is defined as follows:

$$I_n^*(\lambda) = \widehat{q}(\lambda) I_n^+(\lambda), \ \lambda \in [0, \pi].$$

Under some standard assumptions, the validity of this hybrid bootstrap was shown in Kreiss and Paparoditis (2003) for spectral means (e.g., sample autocovariance and spectral distribution function)

$$\int_0^{\pi} \varphi(\omega) I_n(\omega) d\omega, \tag{17}$$

where it is necessary to fit (at least asymptotically) the correct model and for ratio statistics (e.g., sample autocorrelation)

$$\int_0^{\pi} \varphi(\omega) I_n(\omega) d\omega \Big/ \int_0^{\pi} I_n(\omega) d\omega \tag{18}$$

and kernel spectral estimators, where it is not necessary to fit the correct model.

As can be seen from Kreiss and Paparoditis (2003), the described hybrid bootstrap procedure works well, and indeed the effect that on one hand the nonparametric correction step in frequency domain corrects for features which cannot be represented by the autoregressive model and that on the other hand the superior properties of the

autoregressive bootstrap procedure show up can be observed. Especially, it is observed that the frequency domain part of the described hybrid bootstrap leads to a much less dependence of the hybrid bootstrap on the selected autoregressive order p than for the parametric autoregressive bootstrap itself.

The so far described hybrid bootstrap procedure is applicable to statistics, which can be written as functions of the periodogram only. But of course, relevant statistics in time series analysis do not share this property as, for example, the simple sample mean of the observations. Therefore, one is interested in a resampling procedure which still uses some computational parts in frequency domain but which are able to produce bootstrap observations $X_1^*, \ldots, X_n^*$ in time domain. When we switch to the frequency domain, as is, for example, suggested in Step 3 above, then we have to take into account the fact that the periodogram I_n^+ does not contain all information about the bootstrap process X^+ that is contained in the bootstrap observations $X_1^+, \ldots, X_n^+$. But, we can write $I_n^+(\omega) = |J_n^+(\omega)|^2$, where

$$J_n^+(\omega) = \frac{1}{\sqrt{2\pi n}} \sum_{s=1}^{n} X_s^+ \exp^{-is\omega} \tag{19}$$

denotes the discrete Fourier-transform (DFT). And of course, there is a one-to-one correspondence between the n observations of a time series and the DFT evaluated at the Fourier frequencies $\omega_j = 2\pi \frac{j}{n}$ (cf. (16)). The solution now is to apply a nonparametric correction in the frequency domain to the DFT instead of the periodogram and then use the one-to-one correspondence to get back to the time domain. The modified hybrid bootstrap procedure reads as follows:

Step 1: Fit an $AR(p)$ model to the data, compute the estimated residuals $\widehat{\epsilon}_t = X_t - \sum_{v=1}^{p} \widehat{a}_v(p) X_{t-v}, t = p+1, \ldots, n$.

Step 2: Generate bootstrap observations $X_1^+, \ldots, X_n^+$ according to $X_t^+ = \sum_{v=1}^{p} \widehat{a}_v(p) X_{t-v}^+ + \widehat{\sigma}(p)\epsilon_t^+$, ϵ_t^+ i.i.d. with empirical distribution of standardized residuals.

Step 3: Compute the DFT $J_n^+(\omega)$ and the nonparametric correction term $\widetilde{q}(\omega) = \widehat{q}^{1/2}(\omega)$ at the fourier frequencies $\omega_j = 2\pi \frac{j}{n}$, $j = 1, \ldots, n$.

Step 4: Compute the inverse DFT of the corrected DFT $\widetilde{q}(\omega_1) J_n^+(\omega_1), \ldots, \widetilde{q}(\omega_n) J_n^+(\omega_n)$ to obtain bootstrap observations $X_1^*, \ldots, X_n^*$ according to

$$X_t^* = \sqrt{\frac{2\pi}{n}} \sum_{j=1}^{n} \widetilde{q}(\omega_j) J_n^+(\omega_j) e^{it\omega_j}, \quad t = 1, \ldots, n. \tag{20}$$

This modified hybrid bootstrap proposal works for spectral means and ratio statistics as the not modified hybrid bootstrap procedure of Kreiss and Paparoditis (2003) does. Instead of using representations of statistics in frequency domain, we now simply can compute statistics in the time domain. The paper Jentsch and Kreiss (2010), to which we refer for details, discusses the modified hybrid bootstrap procedure for the multivariate case which in many respects is different.

So far, we only have considered autoregressions as parametric models to which we apply nonparametric corrections in frequency domain. It is of course not necessary

that the underlying model follows an autoregressive scheme of finite or infinite order, because of the additional nonparametric correction step. Moreover, it is not necessary to stay with autoregressive models; this has been done for simplicity only. So concerning a hybrid bootstrap procedure, one may think of any parametric model fit in a first step and a nonparametric correction as has been described in a second step. In the univariate situation, the resulting hybrid bootstrap procedure will result in asymptotically correct approximation results for statistics of observations from linear processes, which can be written as functions of autocorrelations or of the standardized (having integral one) spectral density as well as typically for the sample mean. The main reason for that is that asymptotic distributions of such statistics only depend on second-order terms of the underlying stochastic process, and these quantities are correctly mimicked by a hybrid bootstrap proposals. In the multivariate case, the mentioned result concerning the dependence of asymptotic distribution on second-order terms of linear time series does not hold any more, and therefore, the multivariate situation is much more involved (cf. Jentsch and Kreiss (2010)). A related method that allows resampling in frequency domain to obtain bootstrap replicates in time domain is considered in Kirch and Politis (2011). The papers Sergides and Paparoditis (2008) and Kreiss and Paparoditis (2011) considered an autoregressive-aided frequency domain hybrid bootstrap procedure and the modified hybrid bootstrap procedure along the lines described in this section for locally stationary time series.

8. Bootstrap under long-range dependence

Let $\{X_t\}_{t\in\mathbb{N}}$ be a stationary process with $EX_1^2 \in (0,\infty)$, autocovariance function $r(\cdot)$, and spectral density function $f(\cdot)$. We say that the process $\{X_t\}_{t\in\mathbb{N}}$ is long-range dependent (LRD) if $\sum_{k=1}^{\infty} |r(k)| = \infty$ or if $f(\lambda) \to \infty$ as $\lambda \to 0$. Otherwise, $\{X_t\}_{t\in\mathbb{N}}$ is said to be short-range dependent (SRD). We also use the acronym LRD (SRD) for long- (respectively, short) range dependence. Limit behaviors of many common statistics and tests under LRD are different from their behaviors under SRD. For example, the sample mean of n observations from a LRD process may converge to the population mean at a rate *slower* than $O_p(n^{-1/2})$, and similarly, with proper centering and scaling, the sample mean may have a *non-normal* limit distribution even when the population variance is finite. More specifically, we consider the following result on the sample mean under LRD. Let $\{Z_t\}_{t\in\mathbb{N}}$ be a zero mean unit variance Gaussian process with an autocovariance function $r_1(\cdot)$ satisfying

$$r_1(k) \sim Ck^{-\alpha} \quad \text{as} \quad k \to \infty, \tag{21}$$

for some $\alpha \in (0,1)$, where for any two sequences $\{s_n\}_{n\geq 1}$ in $\mathbb{R}$ and $\{t_n\}_{n\geq 1}$ in $(0,\infty)$, we write $s_n \sim t_n$ if $s_n/t_n \to 1$ as $n \to \infty$. Note that here $\sum_{k=1}^{\infty} |r_1(k)| = \infty$, and hence, the process $\{Z_t\}$ is LRD. Next suppose that the X_t process derives from the Z_t process through the transformation

$$X_t = H_q(Z_t), \quad t \in \mathbb{N}, \tag{22}$$

for some integer $q \geq 1$, where $H_q(x)$ is the qth Hermite polynomial, i.e., for $x \in \mathbb{R}$, $H_q(x) = (-1)^q \big(\exp(x^2/2)\big) \frac{d^q}{dx^q} \big(\exp(-x^2/2)\big)$. Results in Taqqu (1975, 1979) and Dobrushin and Major (1979) imply the following result on the sample mean:

THEOREM 1. *Suppose that $\{X_t\}_{t\in\mathbb{N}}$ admits the representation (22) for some $q \geq 1$. If $\alpha \in (0, q^{-1})$, then*

$$n^{q\alpha/2}(\bar{X}_n - \mu) \to^d W_q \tag{23}$$

where $\mu = EX_1$ and where W_q is defined in terms of a multiple Wiener-Ito integral with respect to the random spectral measure W of the Gaussian white noise process as

$$W_q = A^{-q/2} \int \frac{\exp(\iota(x_1 + \cdots + x_q)) - 1}{\iota(x_1 + \cdots + x_q)} \prod_{k=1}^{q} |x_k|^{(\alpha-1)/2} dW(x_1) \ldots dW_q(x_q) \tag{24}$$

with $A = 2\Gamma(\alpha)\cos(\alpha\pi/2)$.

For $q = 1$, W_q has a normal distribution with mean zero and variance $2/[(1-\alpha)(2-\alpha)]$. However, for $q \geq 2$, W_q has a non-normal distribution. Although the bootstrap methods described in the earlier sections are successful in a variety of problems under SRD, they need not provide a valid answer under LRD. The following result gives the behavior of the MBB approximation under LRD:

THEOREM 2. *Let $\bar{X}_n^*$ denote the MBB sample mean based on blocks of size ℓ and resample size n. Suppose that the conditions of Theorem 1 hold and that $n^\delta \ell^{-1} + \ell n^{1-\delta} = o(1)$ as $n \to \infty$ for some $\delta \in (0,1)$. Then,*

$$\sup_{x\in\mathbb{R}} \Big| P_* \Big(c_n(\bar{X}_n^* - \hat{\mu}) \leq x \Big) - P\Big(n^{q\alpha/2}(\bar{X}_n - \mu) \leq x \Big) \Big| = o(1) \quad as \quad n \to \infty \tag{25}$$

for some sequence $\{c_n\}_{n\geq 1} \in (0,\infty)$ if and only if $q = 1$.

Theorem 2 is a consequence of the results in Lahiri (1993). It shows that for any choice of the scaling sequence, the MBB method fails to capture the distribution of the sample mean whenever the limit distribution of $\bar{X}_n$ is non-normal. With minor modifications of the arguments in Lahiri (1993), it can be shown that the same conclusion also holds for the NBB and the CBB. Intuitively, this may not be very surprising. The heuristic arguments behind the construction of these block bootstrap methods show (cf. Section 5) that all three methods attempt to estimate the initial approximation P_ℓ^∞ to the joint distribution P of $\{X_t\}_{t\in\mathbb{N}}$, but P_ℓ^∞ itself gives an inadequate approximation to P under LRD. Indeed, for the same reason, the MBB approximation fails even for $q = 1$ with the natural choice of the scaling sequence $c_n = n^{q\alpha/2}$. In this case, the (limit) distribution can be captured by using the MBB only with specially constructed scaling sequences $\{c_n\}_{n\geq 1}$, where $c_n \sim [n/\ell^{1+q\alpha}]^{1/2}$ as $n \to \infty$. For the sample mean of an LRD *linear* process with a normal limit, Kim and Nordman (2011) recently established the validity of MBB. Formulation of a suitable bootstrap method that works for both

normal and non-normal cases is still an open problem. For related results on subsampling and empirical likelihood methods under LRD, see Hall et al. (1998), Nordman et al. (2007), and the references therein.

Acknowledgment

The research is partially supported by US NSF grant number DMS 1007703.

References

Athreya, K.B., Fuh, C.D., 1992. Bootstrapping markov chains: countable case. J. Stat. Plan. Inference 33, 311–331.

Athreya, K.B., Ney, P., 1978. A new approach of the limit theory of recurrent Markov chains. Trans. Am. Math. Soc. 245, 493–501.

Basawa, I.V., Mallik, A.K., McCormick, W.P., Reeves, J.H., Taylor, R.L., 1991. Bootstrapping unstable first-order autoregressive processes. Ann. Stat. 19, 1098–1101.

Berkowitz, J., Kilian, L., 2000. Recent developments in bootstrapping time series. Econom. Rev. 19, 1–48.

Bertail, P., Clemencon, S., 2006. Regenerative block bootstrap for Markov chains. Bernoulli 12, 689–712.

Bose, A., 1988. Edgeworth correction by bootstrap in autoregressions. Ann. Stat. 16, 1709–1722.

Bose, A., 1990. Bootstrap in moving average models. Ann. Inst. Stat. Math. 42, 753–768.

Bose, A., Politis, D.N., 1995. A review of the bootstrap for dependent samples. In: Bhat, B.R., Prakasa Rao, B.L.S. (Eds.), Stochastic Processes and Statistical Inference. New Age International Publishers, New Delhi, pp. 39–51.

Brockwell, P., Davis, R.A. (Eds.), 1991. Time Series: Theory and Methods, Springer, New York.

Bühlmann, P., 1997. Sieve bootstrap for time series. Bernoulli 3, 123–148.

Bühlmann, P., 1998. Sieve bootstrap for smoothing in nonstationary time series. Ann. Stat. 26, 48–83.

Bühlmann, P., 2002. Bootstraps for time series. Stat. Sci. 17, 52–72.

Bühlmann, P., Künsch, H.R., 1995. The blockwise bootstrap for general parameters of a stationary time series. Scand. J. Statist. 22, 35–54.

Bühlmann, P., Künsch, H.R., 1999. Block length selection in the bootstrap for time series. Comput. Stat. Data Anal. 31, 295–310.

Carey, V.J., 2005. Resampling methods for dependent data. J. Amer. Statist. Assoc. 100, 712–713.

Carlstein, E., 1986. The use of subseries values for estimating the variance of a general statistics from a stationary time series. Ann. Statist. 14, 1171–1179.

Carlstein, E., Do, K.-A., Hall, P., Hesterberg, T., Künsch, H.R., 1998. Matched-block bootstrap for dependent data. Bernoulli 4, 305–328.

Choi, E., Hall, P., 2000. Bootstrap confidence regions computed from autoregressions of arbitrary order. J. R. Stat. Soc. Ser. B 62, 461–477.

Dahlhaus, R., Janas, D., 1996. A frequency domain bootstrap for ratio statistics in time series analysis. Ann. Statist. 24, 1934–1963.

Datta, S., 1996. On asymptotic properties of bootstrap for AR(1) processes. J. Stat. Plan. Infer 53, 361–374.

Datta, S., McCormick, W.P., 1995a. Bootstrap inference for a first-order autoregression with positive innovations. J. Am. Statist. Assoc. 90, 1289–1300.

Datta, S., McCormick, W.P., 1995b. Some continuous edgeworth expansions for markov chains with application to bootstrap. J. Multivar. Anal. 52, 83–106.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Application, Cambridge University Press, Cambridge, UK.

Dette, H., Paparoditis, E., 2009. Bootstrapping frequency domain tests in multivariate time series with an application to testing equality of spectral densities. J. R. Stat. Soc. Ser. B 71, 831–857.

Dobrushin, R.L., Major, P., 1979. Non-central limit theorems for non-linear functionals of Gaussian fields. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 50, 27–52.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7, 1–26.

Efron, B., 1992. Jackknife-after-bootstrap standard errors and influence functions (disc. pp. 111–127). J. R. Stat. Soc. Ser. B 54, 83–111.
Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap, Chapman and Hall, New York.
Franke, J., Härdle, W., 1992. On bootstrapping kernel spectral estimates. Ann. Stat. 20, 121–145.
Franke, J., Kreiss, J.-P., 1992. Bootstrapping stationary autoregressive moving-average models. J. Time Ser. Anal. 13, 297–317.
Franke, J., Kreiss, J.-P., Mammen, E., 2002a. Bootstrap of kernel smoothing in nonlinear time series. Bernoulli 8, 1–37.
Franke, J., Kreiss, J.-P., Mammen, E., Neumann, M.H., 2002b. Properties of the nonparametric autoregressive bootstrap. J. Time Ser. Anal. 23, 555–585.
Franke, J., Kreiss, J.-P., Moser, M., 2006. Bootstrap order selection for autoregressive processes. Stat. Decis. 24, 305–325.
Gonçalves, S., Kilian, L., 2007. Asymptotic and bootstrap inference for AR(∞) processes with conditional heteroskedasticity. Econom. Rev. 26, 609–641.
Götze, F., Künsch, H., 1996. Second-order correctness of the blockwise bootstrap for stationary observations. Ann. Statist. 24, 1914–1933.
Hall, P., 1992. The Bootstrap and Edgeworth Expansion, Springer, New York.
Hall, P., Horowitz, J.L., Jing, B.-Y., 1995. On blocking rules for the bootstrap with dependent data. Biometrika 82, 561–574.
Hall, P., Jing, B.-Y., Lahiri, S.N., 1998. On the sampling window method under long range dependence. Stat. Sin. 8, 1189–1204.
Härdle, W., Horowitz, J., Kreiss, J.-P., 2003. Bootstrap for time series. Int. Stat. Rev. 71, 435–459.
Heimann, G., Kreiss, J.-P., 1996. Bootstrapping general first order autoregression. Stat. Prob. Lett. 30, 87–98.
Horowitz, J.L., 2003. Bootstrap methods for markov processes. Econometrica 71, 1049–1082.
Hurvich, C.M., Zeger, S.L., 1987. Frequency Domain Bootstrap Methods for Time Series. Preprint, Department of Statistics and Operations Research, New York University.
Jentsch, C., Kreiss, J.-P., 2010. The multiple hybrid bootstrap: Resampling multivariate linear processes. J. Mult. Anal. 101, 2320–2345.
Kapetanios, G., 2010. A generalization of a sieve bootstrap invariance principle to long memory processes. Quant. Qual. Anal. Soc. Sci. 4, 19–40.
Kim, Y.-M., Nordman, D.J., 2011. Properties of a block bootstrap method under long range dependence. Sankhya Ser. A. 73, 79–109.
Kirch, C., Politis, D.N., 2011. TFT-Bootstrap: Resampling time series in the frequency domain to obtain replicates in the time domain. Ann. Stat. 39, 1427–1470.
Kreiss, J.-P., 1988. Asymptotic Statistical Inference for a Class of Stochastic Processes, Habilitationsschrift, Universität Hamburg.
Kreiss, J.-P., 1992. Bootstrap procedures for AR(∞)-processes. In: Lecture Notes in Economics and Mathematical Systems, vol. 376 (Proc. Bootstrapping and Related Techniques, Trier), pp. 107–113.
Kreiss, J.-P., 1997. Asymptotical Properties of Residual Bootstrap for Autoregression. Preprint, TU Braunschweig.
Kreiss, J.-P., 2000. Nonparametric estimation and bootstrap for financial time series. In: Chan, W.S., Li, W.K., Tong, H., (Eds.), Statistics and Finance: An Interface. Imperial College Press, London.
Kreiss, J.-P., Neuhaus, G., 2006. Einführung in die Zeitreihenanalyse, Springer, Heidelberg.
Kreiss, J.-P., Neumann, M.H., 1999. Bootstrap tests for parametric volatility structure in nonparametric autoregression. In: Grigelionis, B. et al., (Eds.), Prob. Theory Math. Stat. pp. 393–404.
Kreiss, J.-P., Neumann, M.H., Yao, Q., 2008. Bootstrap tests for simple structures in nonparametric time series regression. Stat. Inter. 1, 367–380.
Kreiss, J.-P., Paparoditis, E., 2003. Autoregressive aided periodogram bootstrap for time series. Ann. Stat. 31, 1923–1955.
Kreiss, J.-P., Paparoditis, E., 2011. Bootstrapping Locally Stationary Time Series. Technical Report.
Kreiss, J.-P., Paparoditis, E., Politis, D.N., 2011. On the range of validity of the autoregressive sieve bootstrap. Ann. Stat. 39, 2103–2130.
Kulperger, R.J., Prakasa Rao, B.L.S., 1989. Bootstrapping a finite state Markov chain. Sankhya Ser. A 51, 178–191.
Künsch, H.R., 1989. The jackknife and the bootstrap for general stationary observations. Ann. Statist. 17, 1217–1241.

Lahiri, S.N., 1991. Second order optimality of stationary bootstrap. Stat. Prob. Lett. 11, 335–341.
Lahiri, S.N., 1992. Edgeworth correction by moving block bootstrap for stationary and nonstationary data. In: LePage, R., Billard, L. (Eds.), Exploring the Limits of Bootstrap. Wiley, New York, pp. 183–214.
Lahiri, S.N., 1993. On the moving block bootstrap under long range dependence. Stat. Prob. Lett. 18, 405–413.
Lahiri, S.N., 1996. On edgeworth expansions and the moving block bootstrap for studentized M–estimators in multiple linear regression models. J. Multivar. Anal. 56, 42–59.
Lahiri, S.N., 1999. Theoretical comparisons of block bootstrap methods. Ann. Statist. 27, 386–404.
Lahiri, S.N., 2002. On the jackknife after bootstrap method for dependent data and its consistency properties. Econom. Theory 18, 79–98.
Lahiri, S.N., 2003a. Resampling Methods for Dependent Data, Springer, New York.
Lahiri, S.N., 2003b. A necessary and sufficient condition for asymptotic independence of discrete Fourier transforms under short- and long-range dependence. Ann. Stat. 31, 613–641.
Lahiri, S.N., Furukawa, K., Lee, Y.-D., 2007. A nonparametric plug-in method for selecting the optimal block length. Stat. Method 4, 292–321.
Li, H., Maddala, G.S., 1996. Bootstrapping time series models. Econom. Rev. 15, 115–158.
Liu, R.Y., Singh, K., 1992. Moving blocks jackknife and bootstrap capture weak dependence. In: LePage, R., Billard, L. (Eds.), Exploring the Limits of Bootstrap. Wiley, New York.
Neumann, M. H., Kreiss, J.-P., 1998. Regression-type inference in nonparametric autoregression. Ann. Stat. 26, 1570–1613.
Nordman, D.J., 2009. A note on the stationary bootstrap. Ann. Stat. 37, 359–370.
Nordman, D., Sibbersten, P., Lahiri, S.N., 2007. Empirical likelihood confidence intervals for the mean of a long range dependent process. J. Time Ser. Anal. 28, 576–599.
Paparoditis, E., 1996. Bootstrapping autoregressive and moving average parameter estimates of infinite order vector autoregressive processes. J. Multivar. Anal. 57, 277–296.
Paparoditis, E., 2002. Frequency domain bootstrap for time series. In: Dehling, H., Mikosch, T., Sørensen, M. (Eds.), Empirical Process Techniques for Dependent Data. Birkhäuser, Boston, pp. 365–381.
Paparoditis, E., Politis, D.N., 1999. The local bootstrap for periodogram statistics. J. Time Ser. Anal. 20, 193–222.
Paparoditis, E., Politis, D.N., 2000. The local bootstrap for kernel estimators under general dependence conditions. Ann. Inst. Statist. Math. 52, 139–159.
Paparoditis, E., Politis, D.N., 2001a. The tapered block bootstrap. Biometrika 88, 1105–1119.
Paparoditis, E., Politis, D.N., 2001b. A markovian local resampling scheme for nonparametric estimators in time series analysis. Econom. Theory 17, 540–566.
Paparoditis, E., Politis, D.N., 2003. Residual-based block bootstrap for unit root testing. Econometrica 71, 813–856.
Paparoditis, E., Politis, D.N., 2005. Bootstrapping unit root tests for autoregressive time series. J. Am. Statist. Assoc. 100, 545–553.
Paparoditis, E., Politis, D.N., 2009. Resampling and subsampling for financial time series. In: Andersen, T., Davis, R., Kreiss, J.-P., Mikosch, T. (Eds.), Handbook of Financial Time Series. Springer, New York, pp. 983–999.
Paparoditis, E., Streitberg, B., 1991. Order identification statistics in stationary autoregressive moving average models: vector autoccorrelations and the bootstrap. J. Time Ser. Anal. 13, 415–435.
Park, J.Y., 2002. An invariance principle for siebe bootstrap in time series. Econom. Theory 18, 469–490.
Patton, A., Politis, D.N., White, H., 2009. Correction to "Automatic block-length selection for the dependent bootstrap by D.N. Politis and H. White". Econom. Rev. 28, 372–375.
Politis, D.N., 2003. The impact of bootstrap methods on time series analysis. Stat. Sci. 18, 219–230.
Politis, D.N., Romano, J.P., 1992. A circular block resampling procedure for stationary data. In: Lepage, R., Billard, L., (Eds.), Exploring the Limits of Bootstrap. Wiley, New York, pp. 263–270.
Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. J. Am. Statist. Assoc. 89, 1303–1313.
Politis, D.N., White, H., 2004. Automatic block-length selection for the dependent bootstrap. Econom. Rev. 23, 53–70.
Poskitt, D.S., 2008. Properties of the sieve bootstrap for fractionally integrated and non-invertible processes. J. Time Ser. Anal. 29, 224–250.
Psaradakis, Z., 2001. Bootstrap tests for an autoregressive unit root in the presence of weakly dependent errors. J. Time Ser. Anal. 22, 577–594.

Rajarshi, M.B., 1990. Bootstrap in Markov sequences based on estimates of transition density. Ann. Inst. Statist. Math. 42, 253–268.

Ruiz, E., Pascual, L., 2002. Bootstrapping financial time series. J. Econ. Surveys 16, 271–300. Reprinted in: Contributions to Financial Econometrics: Theoretical and Practical Issues (Eds.: McAleer, M. and Oxley, L.), Blackwell.

Sergides, M., Paparoditis, E., 2008. Bootstrapping the local periodogram of locally stationary processes. J. Time Ser. Anal. 29, 264–299. Corrigendum: Journal of Time Series Analysis 30, 260–261.

Shao, J., Tu, D., 1995. The Jackknife and Bootstrap, Springer, New York.

Singh, K., 1981. On the asymptotic accuracy of Efron's bootstrap. Ann. Stat. 9, 1187–1195.

Taqqu, M.S., 1975. Weak convergence to fractional Brownian motion and to the Rosenblatt process. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 31, 287–302.

Taqqu, M.S., 1979. Convergence of integrated processes of arbitrary hermite rank. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 50, 53–83.

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00002-8

Testing Time Series Linearity: Traditional and Bootstrap Methods

Arthur Berg[1], *Timothy McMurry*[2] *and Dimitris N. Politis*[3]

[1]*Division of Biostatistics, Penn State University, Hershey, PA 17033, USA*
[2]*Division of Biostatistics, University of Virginia, Charlottesville, VA 22908-0717, USA*
[3]*Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112, USA*

Abstract

We review the notion of time series linearity and describe recent advances in linearity and Gaussianity testing via data resampling methodologies. Many advances have been made since the first published tests of linearity and Gaussianity by Subba Rao and Gabr in 1980, including several resampling-based proposals. This chapter is intended to be instructive in explaining and motivating linearity testing. Recent results on the validity of the AR-sieve bootstrap for linearity testing are reviewed. In addition, a subsampling-based linearity and Gaussianity test is proposed where asymptotic consistency of the testing procedure is justified.

Keywords: AR-sieve, asymptotic consistency, bootstrap, gaussianity testing, linearity testing, literature review, subsampling, time series.

1. Introduction

Ever since the fundamental recognition of the potential role of the computer in modern statistics (Efron, 1979a,b), the bootstrap and other resampling methods have been extensively developed for inference in independent data settings; see, e.g., the works done by Davison and Hinkley (1997), Efron and Tibshirani (1993), Hall (1997), Shao and Tu (1995). Such methods are even more important in the context of dependent data where the distribution theory for estimators and test statistics may be difficult to obtain even asymptotically.

In the time series context, different resampling and subsampling methods have been proposed, and are currently receiving the attention of the statistical community. Reviews of the impact of bootstrap methods on time series analysis may be found in books (Lahiri, 2003; Politis et al., 1999), papers (Bühlmann, 2002; Politis, 2003), and the review by J.-P. Kreiss and S. N. Lahiri in this volume of the Handbook.

In the paper at hand, we revisit the problem of assessing whether a given time series is linear versus nonlinear, or Gaussian versus non-Gaussian. In practice, a Gaussian classification would indicate an Autoregressive Moving Average (ARMA) model with Gaussian innovations is appropriate, whereas a linear classification would indicate that an ARMA model with independent but possibly non-Gaussian innovations can still be considered. However, the rejection of linearity typically requires the practitioner to carefully select an appropriate nonlinear model for the underlying time series, or even to proceed in a model-free, nonparametric manner.

We review the traditional linearity and Gaussianity tests that are based on the normalized bispectrum. The critical regions of these tests have been traditionally determined via asymptotic methods. As an alternative, we describe how these critical regions can be determined via resampling (e.g., the AR-sieve bootstrap) and/or subsampling. One of the advantages of subsampling methodology is the generality under which it is valid. There are a number of examples where subsampling yields consistent estimation but the bootstrap fails (Politis et al., 1999). Although subsampling is more widely applicable, it is noted that when the bootstrap is indeed valid it may possess second-order asymptotic properties (Hall, 1997) giving the bootstrap an advantage.

The literature on linearity and Gaussianity tests is reviewed in the next section. The concept of time series linearity is thoroughly described in Section 3. Sections 4 and 5 focus on the AR-sieve bootstrap and subsampling tests, respectively.

2. A brief survey of linearity and Gaussianity tests

Several parametric and semiparametric tests of linearity designed with a specific nonlinear model as an alternative hypothesis have been proposed, including the works of An et al. (2000), Ashley and Patterson (2009), Chan (1990), Chan and Tong (1986, 1990), Hansen (1999), Harvey and Leybourne (2007), Keenan (1985), Luukkonen et al. (1988), Petruccelli (1990), Petruccelli and Davies (1986), Terasvirta (1994), Terasvirta et al. (1993) and, Tsay (1986). Some tests have model-based assumptions on the null hypothesis (e.g., assuming the null to be AR(p) where p may or may not be assumed known) and some tests induce model-based assumptions on the alternative hypothesis (e.g., assuming the specific GARCH nonlinear alternative hypothesis). Such model-based assumptions may help to increase the power of the various tests, but only when the respective assumptions are satisfied.

Many nonparametric or model-free tests, including the first published linearity test done by Subba Rao and Gabr (1980), are based on nonparametric estimates of the normalized bispectrum, and thus involve much less restrictive assumptions under the null and alternative hypothesis; the normalized bispectrum will be defined and discussed in Section 3. Further bispectrum-based tests include the tests done by Ashley et al. (1986), Birkelund and Hanssen (2009), Brockett et al. (1988), Hinich (1982), Jahan and Harvill

(2008), Subba Rao and Gabr (1984), and Yuan (2000). Tests based on the normalized bispectrum are frequently used in practice when data are available in abundance, for example, when analyzing financial time series; see, e.g., the works done by Abhyankar et al. (1995, 1997), Hinich and Patterson (1985, 1989), and Hsieh (1989) . Note that there are other nonparametric or model-free tests of linearity that are not based on the normalized bispectrum; see, e.g., the works done by Hong-Zhi and Bing (1991), Terdik and Math (1998), and Theiler et al. (1992). An overview of some of these tests is provided in the works of Corduas (1994). In this volume, Giannerini (2011) provides an overview of several linearity testing approaches.

Because of the nonparametric nature of the bispectrum-based tests, their critical regions have traditionally been determined via asymptotic approximations. However, considerably large sample sizes can be necessary in order to accurately estimate the two-dimensional bispectral density. As such, a number of resampling-based methods have been proposed in the recent literature to overcome this limitation in a finite sample size setting.

There are many published reports, especially in recent years, that utilize some form of resampled data in linearity testing (Berg et al., 2010; Birkelund and Hanssen, 2009; Hinich et al., 2005; Hjellvik and Tjostheim, 1995; Kugiumtzis, 2008). Many of these methods involve bootstrapping residuals obtained from fitting a parametric model which is equivalent to resampling the data obtained after a prewhitening step that has removed (to large extent) the presence of autocorrelation. If the prewhitening is performed by fitting an autoregressive [AR(p)] model to the data, then typically practitioners would choose the order p in a data-dependent manner, say by minimizing an information criterion such as AIC, BIC, etc. In practice, it is extremely rare that a finite-order AR(p) would explain the data perfectly; more often than not, the practitioner would use an order p that would be an increasing function of the sample size n, thereby creating an approximating *sieve* of AR models. This is the essence of the AR-sieve bootstrap that is reviewed in detail in the chapter by J.-P. Kreiss and S. N. Lahiri in this volume of the Handbook; the application of the AR-sieve bootstrap to linearity testing is discussed in Section 4.

Another popular approach for linearity testing is the *surrogate data* approach of Theiler et al. (1992). The idea of the surrogate data[1] method is to apply the bootstrap on the phases of the Discrete Fourier Transform (DFT) of the data while keeping the magnitudes of the DFT unchanged. With an inverse DFT, bootstrap pseudo-series can then be created. It is immediate that these pseudo-series have identical second-order structure as the original series, since the second-order structure is coded in the periodogram which remains unchanged in this process.

Alternative uses of the bootstrap in the literature of linearity and Gaussianity testing include a phase scrambling bootstrap (Barnett and Wolff, 2005), the use of bootstrapped residuals to obtain the correct false alarm rate (Hinich et al., 2005; Birkelund and Hanssen, 2009), and the Time Frequency Toggle (TFT)-bootstrap (Kirch and Politis, 2011). The TFT-bootstrap can actually be seen as a generalization of the

[1] In this chapter, we reserve the term *surrogate data* for the method of Theiler et al. (1992); however, the reader should be warned that other authors use the term as a generic way of referring to bootstrap data including even the AR-sieve bootstrap (Hinich et al., 2005; Theiler and Prichard, 1997).

surrogate data method since it involves resampling of both the phases and the magnitudes of the Fourier coefficients. Several surrogate and bootstrap tests for linearity in time series were compared by Kugiumtzis (2008). Finally, a different test has recently been proposed that combines an entropy measure of (non)linearity with bootstrap critical regions (Giannerini et al., 2011).

In this chapter, we chose to highlight two resampling-based tests for time series linearity and Gaussianity. The first is the aforementioned AR-sieve method that bootstraps the residuals obtained from an appropriate AR(p) fit. The AR-sieve methodology has been popular for quite some time but its validity for testing Gaussianity or linearity has only recently been proven (Berg et al., 2010); it is discussed in Section 4. In Section 5, we also describe in detail a novel subsampling-based approach to Gaussianity and linearity testing. The next section defines and discusses the notion of linearity in time series.

3. Linear and nonlinear time series

Consider data $X_1, \ldots, X_n$ arising from a strictly stationary time series $\{X_t\}$ that – for ease of notation – is assumed to have mean zero.[2] The most basic tool for quantifying the inherent strength of dependence is given by the autocovariance function $\gamma(k) = EX_tX_{t+k}$ and the corresponding Fourier series $f(w) = (2\pi)^{-1}\sum_{k=-\infty}^{\infty}\gamma(k)e^{-iwk}$; the latter function is termed the *spectral density*, and is well defined (and continuous) when $\sum_k |\gamma(k)| < \infty$. We can also define the autocorrelation function (ACF) as $\rho(k) = \gamma(k)/\gamma(0)$. If $\rho(k) = 0$ for all $k > 0$, then the series $\{X_t\}$ is said to be a *white noise*, i.e., an uncorrelated sequence; the reason for the term 'white' is the constancy of the associated spectral density function.

The function $\gamma(k)$ represents the second-order moments of the time series $\{X_t\}$; more technically, it represents the second-order *cumulants* (Brillinger, 2001; Rosenblatt, 1985). The third-order cumulants are encapsulated by the function $\Gamma(j,k) = EX_tX_{t+j}X_{t+k}$ and the resulting two-dimensional Fourier series

$$f(w_1, w_2) = (2\pi)^{-2}\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\Gamma(j,k)e^{-iw_1j-iw_2k}$$

is termed the *bispectral density*. For reasons to be apparent soon, we also define the *normalized bispectrum* as

$$K(w_1, w_2) = \frac{|f(w_1, w_2)|^2}{f(w_1)f(w_2)f(w_1+w_2)}.$$

We can similarly define the cumulants of higher order whose corresponding multidimensional Fourier series are termed higher order spectral densities or *polyspectra*; see Section 5 for details. The set of cumulant functions of *all* orders, or equivalently the set of *all* higher order spectral density functions, is a complete description

[2] Centering the data at their sample mean (instead of the true mean) is perfectly acceptable for the subsequent discussion as the resulting error is negligible.

of the dependence structure of the general time series $\{X_t\}$. Of course, working with an infinity of functions is intractable; a welcome shortcut is offered by the notion of linearity.

A time series $\{X_t\}$ is called *linear* if it satisfies an equation of the type:

$$X_t = \sum_{k=-\infty}^{\infty} \beta_k Z_{t-k}, \tag{1}$$

where the coefficients β_k are (at least) square-summable, and the series $\{Z_t\}$ is independent, identically distributed (i.i.d.) with mean zero and variance $\sigma^2 > 0$. To avoid the confounding of the β's with the scale parameter σ, it is helpful to assume that $\beta_0 = 1$.

A linear time series $\{X_t\}$ is called *causal* if $\beta_k = 0$ for $k < 0$, i.e., if

$$X_t = \sum_{k=0}^{\infty} \beta_k Z_{t-k}. \tag{2}$$

Equation (2) should not be confused with the Wold decomposition that *all* purely nondeterministic time series possess (Hannan and Deistler, 1988). In the Wold decomposition, the "error" series $\{Z_t\}$ is only assumed to be a white noise and not i.i.d.; the latter assumption is much stronger. The causality assumption has been used successfully in the context of nonlinear time series as well; see, e.g., the works of Gourieroux and Jasiak (2005) and Wu (2005).

Linear time series are easy objects to work with since their dependence structure is perfectly captured by the sequence of coefficients $\{\beta_k\}$. To elaborate, if $\{X_t\}$ satisfies Eq. (1), then its autocovariance and spectral density functions are given by $\gamma(k) = \sigma^2 \sum_{s=-\infty}^{\infty} \beta_s \beta_{s+k}$ and $f(w) = (2\pi)^{-1}\sigma^2 |\beta(w)|^2$, respectively; here $\beta(w)$ is the Fourier series of the β_k coefficients, i.e., $\beta(w) = \sum_{k=-\infty}^{\infty} \beta_k e^{iwk}$. In addition, the bispectral density is simply given by

$$f(w_1, w_2) = (2\pi)^{-2} \mu_3 \, \beta(-w_1)\beta(-w_2)\beta(w_1 + w_2), \tag{3}$$

where $\mu_3 = EZ_t^3$ is the third moment of the errors. Similarly, all higher order spectra can be calculated in terms of $\beta(w)$.

It is now apparent that the normalized bispectrum $K(w_1, w_2)$ satisfies:

$$K(w_1, w_2) = \frac{|f(w_1, w_2)|^2}{f(w_1) f(w_2) f(w_1 + w_2)} \overset{\text{linearity}}{=} \frac{(\mu_3)^2}{(2\pi)^2 \sigma^6} \overset{\text{Gaussianity}}{=} 0.$$

As indicated by the right-hand side of the above equation, when the time series is in fact linear, the normalized bispectrum will be constant. Furthermore, if the time series is Gaussian (and therefore also linear), the normalized bispectrum will be constantly equal to zero. These two observations form the basis for a host of test of linearity and/or Gaussianity starting with the original paper of Subba Rao and Gabr (1980). Note, however, that although linearity implies the normalized bispectrum is constant, the converse is not necessarily true. Thus there is the implicit, though presumably unlikely, limitation in producing a falsely negative result in the presence of certain nonlinear or non-Gaussian processes.

A prime example of a linear time series is given by the autoregressive (AR) family in which the time series $\{X_t\}$ satisfies a linear relationship with respect to its own lagged values, namely

$$X_t = \sum_{k=1}^{p} \theta_k X_{t-k} + Z_t \tag{4}$$

with the error process $\{Z_t\}$ being i.i.d. $(0, \sigma^2)$ as in Eq. (1). AR modeling lends itself ideally to the problem of predicting future values of the time series; this is particularly true if the AR model is causal. Causality of an AR model is ensured if all roots of the characteristic polynomial $1 - \sum_{k=1}^{p} \theta_k z^k$ have modulus greater than one; see, e.g., the works of Brockwell and Davis (2009).

For example, let $\hat{X}_{n+1}$ denote the predictor of X_{n+1} on the basis of the observed data $X_1, \ldots, X_n$. It is well known (Billingsley, 1995) that the optimal predictor with respect to Mean Squared Error is given by the conditional expectation, i.e., $\hat{X}_{n+1} = E(X_{n+1}|X_1, \ldots, X_n)$. Thus, $\hat{X}_{n+1} = g_n(X_1, \ldots, X_n)$ where $g_n(\cdot)$ is a (generally nonlinear) function of the data $X_1, \ldots, X_n$. In the case of a *causal* AR model, however, it is easy to show that the function $g_n(\cdot)$ is actually *linear*, and that $\hat{X}_{n+1} = \sum_{k=1}^{p} \theta_k X_{n+1-k}$. Note also the property of "finite memory" in that the prediction function $g_n(\cdot)$ is only sensitive to its last p arguments. Although the finite memory property is specific to finite-order causal AR (and Markov) models, the linearity of the optimal prediction function $g_n(\cdot)$ is a property shared by all *causal* linear time series satisfying Eq. (2); this broad class includes all causal and invertible, i.e., "minimum-phase" (Rosenblatt, 2000), ARMA models with i.i.d. innovations.

However, the property of linearity of the optimal prediction function $g_n(\cdot)$ is shared by a larger class of processes. To define this class, consider a weaker form of (2) that amounts to relaxing the i.i.d. assumption on the errors to the assumption of a martingale difference, i.e., to assume that

$$X_t = \sum_{i=0}^{\infty} \beta_i \nu_{t-i}, \tag{5}$$

where $\{\nu_t\}$ is a stationary martingale difference adapted to $\mathcal{F}_t$, the σ-field generated by $\{X_s, s \le t\}$, i.e., that

$$E[\nu_t|\mathcal{F}_{t-1}] = 0 \text{ and } E[\nu_t^2|\mathcal{F}_{t-1}] = 1 \text{ for all } t. \tag{6}$$

As in the study by Kokoszka and Politis (2011), we will use the term *weakly linear* for a time series $\{X_t\}$ that satisfies Eqs. (5) and (6). As it turns out, the linearity of the optimal prediction function $g_n(\cdot)$ is shared by *all* members of the family of weakly linear time series;[3] see, e.g., Theorem 1.4.2 of Hannan and Deistler (1988).

[3] Nonetheless, the class of time series for which the best predictor is linear is larger than the family of weakly linear series. A prime example of a nonweakly linear time series that actually admits a linear optimal predictor can be given by a series of *squared* financial returns, i.e.,when the series $\{X_t\}$ satisfies $X_t = r_t^2$ for all t, and $\{r_t\}$ is modeled by an ARCH/GARCH model; see the works done by Kokoszka and Politis (2011) for details.

The family of Gaussian sequences is an interesting subset of the class of linear time series. Gaussian series occur when the series $\{Z_t\}$ of Eq. (1) is i.i.d. $N(0,1)$, and they too exhibit the useful linearity of the optimal prediction function $g_n(\cdot)$. To see this, recall that the conditional expectation $E(X_{n+1}|X_1,\ldots,X_n)$ turns out to be a linear function of $X_1,\ldots,X_n$ when the variables $X_1,\ldots,X_{n+1}$ are jointly normal (Brockwell and Davis, 2009).

Furthermore, in the Gaussian case all spectra of order higher than two are identically zero. It follows that all dependence information is concentrated in the spectral density $f(w)$. Thus, the investigation of the dependence structure of a Gaussian series can focus on the simple study of second-order properties, namely the ACF $\rho(k)$ and/or the spectral density $f(w)$. For example, an uncorrelated Gaussian series, i.e., one satisfying $\rho(k) = 0$ for all k, necessarily consists of independent random variables. Note that to check/test whether an estimated ACF, denoted by $\hat{\rho}(k)$, is significantly different from zero, the Bartlett confidence limits are typically used. Bartlett's formula, however, is only valid for linear or weakly linear time series (Francq and Zakoïan, 2009; Hannan and Deistler, 1988; Romano and Thombs, 1996). In the (potentially) nonlinear case, even testing the simple null hypothesis $\rho(1) = 0$ becomes highly nontrivial, and is greatly facilitated by a computer intensive methods such as resampling or subsampling (Politis, 2003, Romano and Thombs, 1996).

4. AR-sieve bootstrap tests of linearity

The popular AR-sieve bootstrap method has also been recently shown to be an effective and robust method for Gaussianity and linearity testing. The following gives the general AR-sieve bootstrap algorithm, including separate procedures for Gaussianity and linearity testing as well as third possibility that sits between Gaussianity and linearity – a linear process with symmetric (though possibly non-Gaussian) innovations. The proof of asymptotic consistency of this procedure – under both the null and the alternative hypotheses – can be found in the study by Berg et al. (2010) along with simulations demonstrating its finite sample effectiveness.

AR-sieve bootstrap Algorithm

Step 0: According to some criterion (AIC, BIC, etc.), choose the order p of the AR(p) model to fit to the data $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$.

Step 1: Fit an AR(p) model to $\{X_t\}$ with estimated coefficients $\hat{\boldsymbol{\theta}}_p = (\hat{\theta}_{1,p}, \hat{\theta}_{2,p}, \ldots, \hat{\theta}_{p,p})$; i.e., $\hat{\boldsymbol{\theta}}_p$ is an estimator for $\boldsymbol{\theta}_p$ where

$$\boldsymbol{\theta}_p = (\theta_{1,p}, \theta_{2,p}, \ldots, \theta_{p,p}) = \underset{(c_1,\ldots,c_p)}{\arg\min}\, \mathrm{E}\left[\left(X_t - \sum_{j=1}^{p} c_j X_{t-j}\right)^2\right].$$

Step 2: Let $\boldsymbol{X}^* = \{X_1^*, X_2^*, \ldots, X_n^*\}$ be a series of n pseudo-observations generated by

$$X_t^* = \sum_{j=1}^{p} \hat{\theta}_{j,p} X_{t-j}^* + u_t^* \quad (t = -b, -b+1, \ldots, 0, 1, \ldots, n) \tag{7}$$

where $X_t^* := 0$ for $t < -b$; the positive number b denotes the so-called 'burn-in' period to ensure (approximate) stationarity of the bootstrap series.

In (7), the u_t^*'s are iid random variables having mean zero and distribution function F_n which is selected based on the purpose of the analysis. One of three distribution functions can be selected depending on the null hypothesis under consideration:

Linear null ($H_0^{(1)}$): If the null hypothesis states the time series is linear, then set $F_n = F_n^{(1)}$ to be the empirical distribution function of the centered residuals $\hat{u}_t - \overline{u}_n$, where

$$\hat{u}_t = X_t - \sum_{j=1}^{p} \hat{\theta}_{j,p} X_{t-j} \quad (t = p, p+1, \ldots, n)$$

and

$$\overline{u}_n = \frac{1}{n-p} \sum_{t=p+1}^{n} \widehat{u}_t.$$

Linear symmetric null ($H_0^{(2)}$): If the null hypothesis states the time series is linear with a symmetric distribution of errors, then set $F_n = F_n^{(2)}$ to be a symmetrized version of $F_n^{(1)}$ obtained by setting $u_t^* = S_t u_t^+$ with $S_t \overset{\text{iid}}{\sim} \text{unif}\{-1, 1\}$ (the discrete uniform distribution on -1 and 1) and $u_t^+ \sim F_n^{(1)}$.

Gaussian null ($H_0^{(3)}$): If the null hypothesis states the time series is linear with Gaussian errors, then set $F_n = F_n^{(3)} = N(0, \widehat{\sigma}_p^2)$, where

$$\widehat{\sigma}_p^2 = \frac{1}{n-p} \sum_{t=p+1}^{n} (\widehat{u}_t - \overline{u}_n)^2.$$

Step 3: Compute $T(\boldsymbol{X}^*)$ from the bootstrap series $\boldsymbol{X}^*$ where $T(\cdot)$ is the chosen statistic for the null hypothesis of interest. In the next section, examples of such statistics are provided for testing Gaussianity and linearity.

Repeat: Steps 2 and 3 are repeated a large number (say B) of times. The empirical distribution of the B bootstrap pseudo-statistics can then be used to approximate the true distribution of $T(\boldsymbol{X})$ under the null hypothesis thus making the test feasible.

For example, consider the aforementioned tests based on nonparametric estimates of the normalized bispectrum. In testing for linearity, the normalized bispectral estimator is evaluated over a grid of points and the variability of the estimates are quantified by the interquartile range. If the time series is in fact nonlinear, then the normalized bispectrum should exhibit great variability yielding an interquartile range larger than what would have been expected under linearity. Therefore, linearity is rejected for large values of the estimated interquartile range; see Section 5.2 for more details. Traditionally, the threshold of such a test has been determined from the asymptotic distribution

of the test statistic under the null; the AR-sieve bootstrap offers us a nonasymptotic alternative critical value – see the works done by Berg et al., 2010 for details.

In closing, note that a new bootstrap method for time series, the Linear Process Bootstrap (LPB), has been recently introduced (McMurry and Politis, 2010). The LPB generates linear time series in the bootstrap world whether the true model is linear or not, i.e., under the null of linearity but also under the alternative. As in the AR-sieve bootstrap case, this property makes the LPB bootstrap a promising alternative in connection with bootstrapping the test of linearity.

5. Subsampling tests of linearity

The general subsampling methodology for time series approximates the distribution of a statistic by evaluating the statistic on subsampled blocks or contiguous subsets of the original time series. As with any resampling procedure, there are certain assumptions required on the data and the statistic to guarantee convergence; however, the assumptions needed to achieve consistency of subsampling are generally weaker or easier to verify than the assumptions required for bootstrap procedures (Politis et al., 1999).

To fix ideas, we consider in detail two statistics: a linearity test statistic, t_n^L, and a Gaussianity test statistic, t_n^G. These test statistics are derived from estimates of the normalized bispectrum, and they are based on the statistics originally proposed by Hinich (1982). Whereas Hinich utilized asymptotic theory to determine the distribution of the statistics under their respective null hypotheses, the approach described here uses subsampling to approximate the distributions of the statistics.

The test statistics t_n^L and t_n^G are described and the asymptotic conditions needed to justify the subsampling tests based on these statistics are provided. These test statistics are based on estimates of the spectral density and the bispectrum. Therefore, we first present some theory for polyspectral inference followed by the bispectrum-based method of linearity and Gaussianity testing.

5.1. Kernel-based polyspectral estimation

The assumption of sth-order stationarity is required to define the sth-order polyspectrum. It requires that all moments of order $m \leq s$ to exist and be lag-invariant, i.e.,

$$\mathrm{E}\left[X_{\tau_1} X_{\tau_2} \cdots X_{\tau_m}\right] = \mathrm{E}\left[X_{\tau_1+t} X_{\tau_2+t} \cdots X_{\tau_m+t}\right]$$

for any set of integers $\tau_1, \ldots, \tau_m$ and t. This assumption lies between the weaker assumption of covariance-stationarity (same as second-order stationarity and wide sense stationarity) and the stronger assumption of strict stationarity (also known as strong stationarity).

Let $X_1, X_2, \ldots, X_n$ be a realization of an sth-order stationary time series with (possibly nonzero) mean μ. The sth-order joint cumulant is defined as

$$C(\tau_1, \ldots, \tau_{s-1}) = \sum_{(\nu_1, \ldots, \nu_p)} (-1)^{p-1}(p-1)\, \mu_{\nu_1} \cdots \mu_{\nu_p}, \tag{8}$$

where the sum is over all partitions $(\nu_1, \ldots, \nu_p)$ of $\{0, \ldots, \tau_{s-1}\}$ and $\mu_{\nu_j} = E\left[\prod_{\tau_i \in \nu_j} X_{\tau_i}\right]$; refer to the works done by Jammalamadaka et al. (2006) for another expression of the joint cumulant. The sth-order spectral density is defined as

$$f(\boldsymbol{\omega}) = \frac{1}{(2\pi)^{s-1}} \sum_{\boldsymbol{\tau} \in \mathbb{Z}^{s-1}} C(\boldsymbol{\tau}) e^{-i\boldsymbol{\tau}\cdot\boldsymbol{\omega}}, \tag{9}$$

where the bold-face notation $\boldsymbol{\omega}$ denotes an $(s-1)$-dimensional, vector argument, i.e., $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_{s-1})$. We adopt the usual assumption on $C(\boldsymbol{\tau})$ that it be absolutely summable, thus guaranteeing the existence and continuity of the spectral density. A natural estimator of $C(\boldsymbol{\tau})$ is given by

$$\widehat{C}(\tau_1, \ldots, \tau_{s-1}) = \sum_{(\nu_1, \ldots, \nu_p)} (-1)^{p-1}(p-1)!\, \hat{\mu}_{\nu_1} \cdots \hat{\mu}_{\nu_p}, \tag{10}$$

where the sum is overall partitions of $(\nu_1, \ldots, \nu_p)$ of $\{0, \ldots, \tau_{s-1}\}$ and

$$\hat{\mu}_{\nu_j} = \frac{1}{n - \max(\nu_j) + \min(\nu_j)} \sum_{k=-\min(\nu_j)}^{n-\max(\nu_j)} \prod_{t \in \nu_j} X_{t+k}.$$

The previously discussed second- and third-order cumulant functions, as given by $s = 2$ and $s = 3$ in (8), simplify to the following centered expectations:

$$\begin{aligned} C(\tau_1) &= \mathrm{E}\left[(X_t - \mu)(X_{t+\tau_1} - \mu)\right] \\ C(\tau_1, \tau_2) &= \mathrm{E}\left[(X_t - \mu)(X_{t+\tau_1} - \mu)(X_{t+\tau_2} - \mu)\right]. \end{aligned}$$

In these cases, the corresponding estimator in (10) simplifies to

$$\widehat{C}(\boldsymbol{\tau}) = \frac{1}{n} \sum_{t=1}^{n-\gamma} \prod_{j=1}^{s} (X_{t-\alpha+\tau_j} - \bar{X}), \tag{11}$$

where $\alpha = \min(0, \tau_1, \ldots, \tau_{s-1})$ and $\gamma = \max(0, \tau_1, \ldots, \tau_{s-1}) - \alpha$, and $\bar{X}$ represents the sample mean of the data. We extend the domain of $\widehat{C}$ to all of $\mathbb{Z}^s$ by defining $\widehat{C}(\boldsymbol{\tau}) = 0$ when the sum in (10) or (11) is empty.

Consistent estimation of the polyspectra (9) is obtained by taking the Fourier transform of the sample cumulant function, $\widehat{C}(\boldsymbol{\tau})$, multiplied by a smoothing kernel κ_m with bandwidth $m = m(n)$ that grows asymptotically with n but with $m/n \to 0$; in other words, let

$$\hat{f}(\boldsymbol{\omega}) = \frac{1}{(2\pi)^{s-1}} \sum_{\|\boldsymbol{\tau}\| < n} \kappa_m(\boldsymbol{\tau}) \widehat{C}(\boldsymbol{\tau}) e^{-i\boldsymbol{\tau}\cdot\boldsymbol{\omega}}. \tag{12}$$

Typically, the kernel κ_m is obtained by "dilation" of a fixed underlying kernel κ, i.e., letting $\kappa_m(\boldsymbol{\tau}) = \kappa(\boldsymbol{\tau}/m)$. Several different shapes for κ have been proposed in the

literature, particularly for second-order spectral density estimation; cf. the study by Priestley (1983). In particular, utilizing a "flat-top" lag-window function, such as the trapezoidal function (Politis and Romano, 1995) or the conical frustum (Politis, 2011), will yield a (poly)spectral density estimate with optimal mean square error properties.

Asymptotic theory of the kernel-based polyspectral density estimators (12) is detailed in the works done by Berg and Politis (2009), Brillinger and Rosenblatt (1967), and Rosenblatt (1985). Two assumptions are generally required:

ASSUMPTION 1. *The cumulant function* $C(\tau_1, \ldots, \tau_{s-1})$ *satisfies*

$$\sum_{(t_1,\ldots,t_{s-1})\in\mathbb{Z}^{s-1}} t_j C(t_1, \ldots, t_{s-1}) \qquad \text{for each } j = 1, \ldots, s-1.$$

This assumption implies the existence of a continuously differentiable polyspectral density.

ASSUMPTION 2. *The kernel* $\kappa(\boldsymbol{\tau})$ *is continuously differentiable and satisfies*

$$\max\left(|\tau_j \kappa(\boldsymbol{\tau})|, \left|\frac{\partial}{\partial \tau_j}\kappa(\boldsymbol{\tau})\right|\right) \le M(1 + \|\boldsymbol{\tau}\|)^{-(s-1)-\epsilon} \qquad \text{for each } j = 1, \ldots, s-1,$$

where $\|\boldsymbol{\tau}\| = \left(\sum_{j=1}^{s-1} \tau_j^2\right)^{1/2}$, $M > 0$, *and* $\epsilon > 0$.

If $\{X_t\}$ is a strictly stationary process, Assumptions 1 and 2 can be used to show that

$$\sqrt{n/m^{s-1}}\left(\hat{f}(\boldsymbol{\omega}) - \mathrm{E}\left[\hat{f}(\boldsymbol{\omega})\right]\right) \longrightarrow_d \mathcal{N}\left(0, \sigma^2\right) \tag{13}$$

when $n \to \infty$ but $n/m^{s-1} \to \infty$; here σ^2 is a complex-valued functional of f and κ.

REMARK 1. If the bias of $\hat{f}(\boldsymbol{\omega})$ is of smaller order than $\sqrt{n/m^{s-1}}$, then $\mathrm{E}[\hat{f}(\boldsymbol{\omega})]$ in (13) can be replaced with $f(\boldsymbol{\omega})$. This minimal bias property can be achieved in two ways: (1) by selecting a bandwidth m that is (slightly) bigger than the optimal one resulting in a certain *under*smoothing, or (2) by using an infinite-order kernel κ, which possesses reduced bias properties (Politis, 2011). Selecting an optimal bandwidth in finite samples is an unavoidable issue in nonparametric function estimation; a practical and effective method for selecting an appropriate bandwith for polyspectral estimation is given in the study by Berg and Politis (2009). □

5.2. *The test statistics* t_n^G *and* t_n^L

Due to the symmetries inherent to polyspectra (Berg, 2008), the normalized bispectrum, $K(\omega_1, \omega_2)$ is uniquely defined by its values on Ω given by

$$\Omega := \{(\omega_1, \omega_2) : 0 < \omega_1 < \pi,\ 0 < \omega_2 < \min(\omega_1, 2(\pi - \omega_1))\}.$$

Utilizing estimates of the polyspectra in (12) yields $\hat{K}(\omega_1, \omega_2)$, the estimator of the normalized bispectrum. The Subba Rao and Gabr (1984) Gaussianity test statistic is then defined as

$$t_n^G = \sum_{j=1}^{k} \hat{K}(\omega_j^1, \omega_j^2), \tag{14}$$

where (ω_j^1, ω_j^2) $(j = 1, \ldots, k)$ constitutes a grid of k points inside Ω; the number of points k increases with n to ensure consistency of the test. The null hypothesis of Gaussianity is rejected if t_n^G is too large.

Hinich (1982) proposed an improved and more robust version of the original bispectrum-based linearity test proposed by Subba Rao and Gabr. The Hinich linearity test statistic is given as

$$t_n^L = IQR\left\{\left[\hat{K}(\omega_j^1, \omega_j^2)\right]_{j=1}^{k}\right\}, \tag{15}$$

where IQR stands for the interquartile range. The null hypothesis of linearity is rejected if t_n^L is too large.

In either case, t_n^G or t_n^L, the practitioner must determine the threshold of the critical region, i.e., decide what constitutes "too large" a value of the test statistic. This has been traditionally accomplished via asymptotic arguments (Hinich, 1982, Subba Rao and Gabr, 1984). However, as discussed in Section 4, we can alternatively determine the threshold by a resampling approximation offered by the AR-sieve bootstrap. The following Section describes how to obtain a subsampling approximation to such a critical region.

5.3. *Subsampling for t_n^G and t_n^L*

In order to establish the consistency of subsampling for the test statistics t_n^G and t_n^L, it must be shown that their sampling distribution converges to a continuous limit law under their respective null hypothesis. The asymptotics of the t_n^G and t_n^L have been established in the literature as presented below.

If the time series is Gaussian, then (Subba Rao and Gabr, 1980, 1984)

$$\left(\frac{n}{m^2} \cdot \frac{2\pi}{\zeta_2}\right) t_n^G \longrightarrow_d \chi_{2k}^2, \tag{16}$$

where $m = m(n)$ is the bandwidth used for the estimator (12) and $\zeta_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \kappa^2(\tau_1, \tau_2)\, d\tau_1\, d\tau_2$.

If the time series is linear, then (Berg et al., 2010, Hinich, 1982)

$$\left(\frac{n}{m^2} \cdot \frac{2\pi}{\zeta_2}\right) t_n^L \longrightarrow_d N\left[\xi_{3/4} - \xi_{1/4}, \frac{1}{16k}\left(\frac{3}{g^2(\xi_{1/4})} + \frac{3}{g^2(\xi_{3/4})} - \frac{2}{g^2(\xi_{1/4})g^2(\xi_{3/4})}\right)\right], \tag{17}$$

where $\xi_{\cdot}$ and $g(\cdot)$ are the quantile and density functions, respectively, of the χ^2_{2k} distribution.

Let t_n denote either t_n^G or t_n^L as appropriate. It is easy to see that t_n satisfies the property $t_n \to 0$ under its respective null and $t_n \to t > 0$ under the alternative; convergence of t_n under the alternative is investigated in the works done by Hinich (1982). Define $t_{n,b,t}$ to be the statistic defined by (14) or (15), whichever appropriate, calculated using only the subsample $\{X_t, X_{t+1}, \ldots, X_{t+b-1}\}$ for $t \in \{1, 2, \ldots, n-b+1\}$.

We now consider two candidate subsampling distributions for subsampling the hypothesis test of Gaussianity or linearity. First we define the uncentered subsampling distribution as presented in the study by Politis et al. (1999),

$$S_{n,b}^U(x) := \frac{1}{n-b+1} \sum_{t=1}^{n-b+1} 1\{\tau_b t_{n,b,t} \le x\}, \tag{18}$$

where[4] $\tau_b = b/m(b)^2$.

Alternatively, a centered version of the above subsampling distribution has been shown to possess improved power in many contexts (Berg et al., 2010). The centered subsampling distribution is given by

$$S_{n,b}^C(x) := \frac{1}{n-b+1} \sum_{t=1}^{n-b+1} 1\{\tau_b (t_{n,b,t} - t_n) \le x\}. \tag{19}$$

It follows from (16) and (17) that the sampling distribution of $\tau_n t_n$ converges, under the respective null hypothesis, to a continuous limit law with cumulative distribution function denoted by $H(x)$.The consistency of the subsampling method as applied to linearity and Gaussianity testing is now stated; the following theorem follows directly from Theorem 3.5.1 in Politis et al. (1999).

THEOREM 1 (Validity of subsampling for t_n^G and t_n^L). *Let $H_{n,b}(x)$ denote either $S_{n,b}^U(x)$ or $S_{n,b}^C(x)$. Assume either (16) or (17) according to whether T_n denotes t_n^G or t_n^L. Assume $b \to \infty$, $b/n \to 0$ and $\tau_b/\tau_n \to 0$ as $n \to \infty$. Assume the bandwidth m for the polyspectra estimates used in the construction of the test statistics obeys the undersmoothing condition outlined in Remark 1. Further assume the time series $\{X_t\}$ is strictly stationary, and strong mixing. For $\alpha \in (0,1)$, define the two quantities*

$$h_{n,b}(1-\alpha) = \inf\{x : H_{n,b}(x) \ge 1-\alpha\}$$
$$h(1-\alpha) = \inf\{x : H(x) \ge 1-\alpha\}$$

Then under the null hypothesis

i. $h_{n,b}(1-\alpha) \longrightarrow g(1-\alpha)$ *in probability;*
ii. $Prob\{\tau_n t_n > h_{n,b}(1-\alpha)\} \longrightarrow \alpha \quad$ as $n \to \infty$.

[4] Recall that $m = m(n)$; for example, if $m(n) = n^\delta$ for some $\delta \in (0, 1/2)$, then $\tau_b = b/[b^\delta]^2 = b^{1-2\delta}$.

And under the alternative hypothesis,

iii. $Prob\{\tau_n t_n > h_{n,b}(1-\alpha)\} \longrightarrow 1 \quad \text{as } n \to \infty.$

The above theorem shows that both subsampling distributions $S^U_{n,b}(x)$ or $S^C_{n,b}(x)$ yield consistent α-level tests. However, by analogy to other simpler examples (Berg et al., 2010), we expect that the test based on the centered subsampling distribution $S^C_{n,b}(x)$ would be more powerful than the one based on $S^U_{n,b}(x)$, i.e., that the convergence in part (iii) of the Theorem would be faster when $H_{n,b}(x) = S^C_{n,b}(x)$. By the same token, the convergence in part (ii) of the Theorem is expected to be faster when $H_{n,b}(x) = S^U_{n,b}(x)$, i.e., the level of the test would be more accurately achieved with the uncentered subsampling distribution $S^U_{n,b}(x)$.

References

Abhyankar, A., Copeland, L.S., Wong, W., 1995. Nonlinear dynamics in real-time equity market indices: evidence from the United Kingdom. Econ. J. 105(431), 864–880.

Abhyankar, A., Copeland, L.S., Wong, W., 1997. Uncovering nonlinear structure in real-time stock-market indexes: the S&P 500, the DAX, the Nikkei 225, and the FTSE-100. J. Bus. Econ. Stat. 15(1), 1–14.

An, H.Z., Zhu, L.X., Li, R.Z., 2000. A mixed-type test for linearity in time series. J. Stat. Plan. Inference 88, 339–353.

Ashley, R.A., Patterson, D.M., 2009. A Test of the GARCH(1,1) Specification For Daily Stock Returns. Working paper presented at the 17th Society for Nonlinear Dynamics and Econometrics on April 16, 2009.

Ashley, R.A., Patterson, D.M., Hinich, M.J., 1986. A diagnostic test for nonlinear serial dependence in time series fitting errors. J. Time Ser. Anal. 7(3), 165–178.

Barnett, A.G., Wolff, R.C., 2005. A time-domain test for some types of nonlinearity. IEEE Trans. Signal Process. 53(1), 26–33.

Berg, A., 2008. Multivariate lag-windows and group representations. J. Multivariate Anal. 99(10), 2479–2496.

Berg, A., McMurry, T.L., Politis, D.N., 2010. Subsampling p-values. Stat. Probab. Lett. 80(17-18), 1358–1364.

Berg, A., Paparoditis, E., Politis, D.N., 2010. A bootstrap test for time series linearity. J. Stat. Plan. Inference 140(12), 3841–3857.

Berg, A., Politis, D.N., 2009. Higher-order accurate polyspectral estimation with flat-top lag-windows. Ann. Inst. Stat. Math. 61, 1–22.

Billingsley, P., 1995. Probability and Measure. Wiley Series in Probability and Mathematical Statistics. Wiley-Interscience, New York.

Birkelund, Y., Hanssen, A., 2009. Improved bispectrum based tests for Gaussianity and linearity. Signal Process. 89(12), 2537–2546.

Brillinger, D.R., 2001. Time Series: Data Analysis and Theory. Society for Industrial Mathematics, Philadelphia.

Brillinger, D., Rosenblatt, M., 1967. Spectral Analysis of Time Series, chapter Asymptotic theory of kth order spectra in spectral analysis of time series. Wiley, New York.

Brockett, P.L., Hinich, M.J., Patterson, D., 1988. Bispectral-based tests for the detection of gaussianity and linearity in time series. J. Am. Stat. Assoc. 83(403), 657–664.

Brockwell, P.J., Davis, R.A., 2009. Time Series: Theory and Methods. Springer Verlag, New York.

Bühlmann, P., 2002. Bootstraps for time series. Stat. Sci. 17, 52–72.

Chan, K.S., 1990. Testing for threshold autoregression. Ann. Stat. 18(4), 1886–1894.

Chan, W.S., Tong, H., 1986. On tests for non-linearity in time series analysis. J. Forecast. 5(4), 217–228.

Chan, K.S., Tong, H., 1990. On likelihood ratio tests for threshold autoregression. J. R. Stat. Soc. Series B 52(3), 469–476.

Corduas, M., 1994. Nonlinearity tests in time series analysis. Stat. Methods Appl. 3(3), 291–313.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Application. Cambridge University Press, New York.

Efron, B., 1979a. Bootstrap methods: another look at the jackknife. Ann. Stat. 7(1), 1–26.

Efron, B., 1979b. Computers and the theory of statistics: thinking the unthinkable. SIAM Rev. 21(4), 460–480.

Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap, vol. 57. Chapman & Hall/CRC, New York.

Francq, C., Zakoïan, J.M., 2009. Bartlett's formula for a general class of nonlinear processes. J. Time Ser. Anal. 30(4), 449–465.

Giannerini, S., 2011. The quest for nonlinearity in time series, in: Rao, C.R., Subba Rao, T. (Eds.), Handbook of Statistics, Volume 30: Time Series. Elsevier, Amsterdam, Netherlands.

Giannerini, S., Maasoumi, E., Bee Dagum, E., 2011. Entropy testing for nonlinearity in time series. Technical report, Università di Bologna.

Gourieroux, C., Jasiak, J., 2005. Nonlinear innovations and impulse responses with application to VaR sensitivity. Annales d'Economie et de Statistique (78), 1–31.

Hall, P., 1997. The Bootstrap and Edgeworth Expansion. Springer Verlag, New York.

Hannan, E.J., Deistler, M., 1988. The Statistical Theory of Linear Systems. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Hoboken.

Hansen, B.E., 1999. Testing for linearity. J. Econ. Surv. 13(5), 551–576.

Harvey, D.I., Leybourne, S.J., 2007. Testing for time series linearity. Econom. J. 10(1), 149–165.

Hinich, M.J., 1982. Testing for gaussianity and linearity of a stationary time series. J. Time Ser. Anal. 3(3), 169–176.

Hinich, M.J., Mendes, E.M., Stone, L., 2005. Detecting nonlinearity in time series: surrogate and bootstrap approaches. Stud. Nonlin. Dyn. Econom. 9(4), 3.

Hinich, M.J., Patterson, D.M., 1985. Evidence of nonlinearity in daily stock returns. J. Bus. Econ. Stat. 3(1), 69–77.

Hinich, M.J., Patterson, D.M., 1989. Evidence of nonlinearity in the trade-by-trade stock market return generating process in: Barnett, W.A., Geweke, J., Shell, K. (Eds.), Economic Complexity: Chaos, Sunspots, Bubbles and Nonlinearity-International Symposium in Economic Theory and Econometrics. Cambridge University Press, Cambridge, UK, pp. 383–409.

Hjellvik, V., Tjostheim, D., 1995. Nonparametric tests of linearity for time series. Biometrika 82(2), 351–368.

Hong-Zhi, A., Bing, C., 1991. A Kolmogorov-Smirnov type statistic with application to test for nonlinearity in time series. Int. Stat. Rev. 59(3), 287–307.

Hsieh, D.A., 1989. Testing for nonlinear dependence in daily foreign exchange rates. J. Bus. 62(3), 339–368.

Jahan, N., Harvill, J.L., 2008. Bispectral-based goodness-of-fit tests of gaussianity and linearity of stationary time series. Commun. Stat. Theory Methods 37(20), 3216–3227.

Jammalamadaka, S.R., Rao, T.S., Terdik, G., 2006. Higher order cumulants of random vectors and applications to statistical inference and time series. Sankhyā Indian J. Stat. 68(2), 326–356.

Keenan, D.M.R., 1985. A Tukey nonadditivity-type test for time series nonlinearity. Biometrika 72(1), 39–44.

Kirch, C., Politis, D.N., 2011. TFT-bootstrap: resampling time series in the frequency domain to obtain replicates in the time domain. Ann. Stat. 39(3), 1427–1470.

Kokoszka, P.S., Politis, D.N., 2011. Nonlinearity of ARCH and stochastic volatility models and Bartlett's formula. Probab. Math. Stat. 31(1), 47–59.

Kugiumtzis, D., 2008. Evaluation of surrogate and bootstrap tests for nonlinearity in time series. Stud. Nonlin. Dyn. Econom. 12(1), 4.

Lahiri, S.N., 2003. Resampling Methods for Dependent Data. Springer Series in Statistics. Springer, New York.

Luukkonen, R., Saikkonen, P., Terasvirta, T., 1988. Testing linearity against smooth transition autoregressive models. Biometrika 75(3), 491–499.

McMurry, T., Politis, D.N., 2010. Banded and tapered estimates of autocovariance matrices and the linear process bootstrap. J. Time Ser. Anal. 31, 471–482.

Petruccelli, J.D., 1990. A comparison of tests for setar-type non-linearity in time series. J. Forecast. 9(1), 25–36.

Petruccelli, J., Davies, N., 1986. A portmanteau test for self-exciting threshold autoregressive-type nonlinearity in time series. Biometrika 73(3), 687–694.

Politis, D.N., 2003. The impact of bootstrap methods on time series analysis. Stat. Sci. 18(2), 219–230.
Politis, D.N., 2011. Higher-order accurate, positive semidefinite estimation of large-sample covariance and spectral density matrices. Econom. Theory, vol. 27, 703–744.
Politis, D.N., Romano, J.P., 1995. Bias-corrected nonparametric spectral estimation. J. Time Ser. Anal. 16(1), 67–103.
Politis, D.N., Romano, J.P., Wolf, M., 1999. Subsampling. Springer Verlag, New York.
Priestley, M., 1983. Spectral Analysis and Time Series, vols 1 and 2, Academic Press, New York.
Romano, J.P., Thombs, L.A., 1996. Inference for autocorrelations under weak assumptions. J. Am. Stat. Assoc. 91(434), 590–600.
Rosenblatt, M., 1985. Stationary Sequences and Random Fields. Springer, New York.
Rosenblatt, M., 2000. Gaussian and Non-Gaussian Linear Time Series and Random Fields. Springer Verlag, New York.
Shao, J., Tu, D., 1995. The Jackknife and Bootstrap. Springer, New York.
Subba Rao, T., Gabr, M.M., 1980. A test for linearity of stationary time series. J. Time Ser. Anal. 1(2), 145–158.
Subba Rao, T., Gabr, M.M., 1984. An Introduction to Bispectral Analysis and Bilinear Time Series Models, volume 24 of Lecture Notes in Statistics. Springer, New York.
Terasvirta, T., 1994. Testing linearity and modelling nonlinear time series. Kybernetika 30(3), 319–330.
Terasvirta, T., Lin, C.F., Granger, C.W.J., 1993. Power of the neural network linearity test. J. Time Ser. Anal. 14(2), 209–220.
Terdik, G., Math, J., 1998. A new test of linearity of time series based on the bispectrum. J. Time Ser. Anal. 19(6), 737–753.
Theiler, J., Galdrikian, B., Longtin, A., Eubank, S., Farmer, J.D., 1992. Testing for nonlinearity in time series: the method of surrogate data. Physica D 58, 77–94.
Theiler, J., Prichard, D., 1997. Using 'surrogate surrogate data' to calibrate the actual rate of false positives in tests for nonlinearity in time series. Fields Inst. Comm. 11, 99–113.
Tsay, R.S., 1986. Nonlinearity tests for time series. Biometrika 73(2), 461–466.
Wu, W.B., 2005. Nonlinear system theory: another look at dependence. Proc. Natl. Acad. Sci. U.S.A. 102(40), 14150.
Yuan, J., 2000. Testing linearity for stationary time series using the sample interquartile range. J. Time Ser. Anal. 21(6), 713–722.

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00003-X

The Quest for Nonlinearity in Time Series

Simone Giannerini
Dipartimento di Scienze Statistiche, Università di Bologna,
Via Belle Arti 41, 40126 Bologna, Italy

Abstract

In this chapter, we review the problem of testing for nonlinearity in time series. First, we discuss the definition and the properties of linear processes and the implications that such properties have on the operational strand. Then, we present and review a tentative classification of the various tests that can be found both in the time series and in the nonlinear dynamics literature. Two main factors contributed to the production of a plethora of alternatives for assessing nonlinearity in time series: the first factor is the intrinsic asymmetry between the linear and the nonlinear realm. In fact, there can be departures from linearity in various directions as nonlinear phenomena possess a virtually infinite richness of features. Among such features we can mention irreversibility, nonuniform predictability, noise amplification/suppression, phase synchronization, noise-induced phenomena, sensitivity to initial conditions, and so on. The second factor is the multidisciplinary nature of the problem. Indeed, the problem of characterizing the various aspects of nonlinear processes is shared among different disciplines, such as Statistics, Econometrics, Nonlinear Dynamics, Biology, and Engineering. The review is by no means exhaustive and reflects the personal inclinations of the author.

Keywords: test, nonlinearity, linear prediction, chaos, higher order moments, bispectrum, initial value sensitivity, surrogate data, nonparametric tests, specification tests.

1. Introduction

The linear (Gaussian) paradigm (Box and Jenkins, 1970) provides powerful tools and a simple mathematical framework for analyzing time series data and interpreting phenomena. In many instances, though, its application is either not recommended or

fails to capture essential aspects of the process under study. For instance, it is well known that the business cycle exhibits peculiar asymmetries (Milas et al., 2006); also, volatility, heavy tails, microstructure noise, and irreversibility are the features often associated to financial series. Again, nonuniform predictability, initial value sensitivity, threshold effects, jump phenomena, and multimodality can be found in many fields such as Hydrology, Physics, Biology, Medicine, and so on. Most of the important early work on nonlinearity can be associated to dynamical system theory where complex phenomena are described in terms of deterministic models. It was Henri Poincaré, with his pioneering work on planetary dynamics at the end of the nineteenth century, who first identified the phenomenon of sensitivity to initial conditions. Since then, a number of researchers of different disciplines contributed to the discovery of peculiar features of nonlinear phenomena. For instance, the article on atmospheric dynamics by Lorenz (1963) is usually indicated as the work that popularized the notion of chaos. Other important contributions include the works that laid the foundations of catastrophe theory (Thom, 1989) and fractal geometry (Mandelbrot, 1982).

On the time series analysis strand, Moran (1953) is indicated as one of the first authors who highlighted the limitations of linear models. In his analysis of the Canadian lynx series, Moran observed an "anomaly" in the residuals of a fitted linear model. Such feature was a byproduct of the *regime effect* of the population dynamics; the endeavours to model it contributed to the introduction of the so-called threshold models; see Tong (1990, 2011) and references therein. Again, the need of going beyond second-order moments led to the introduction of bilinear models and higher order spectra (Granger and Andersen, 1978, and Subba Rao and Gabr, 1984). In the study by Subba Rao and Gabr (1984), the authors also observed that the fit and the prediction performance of nonlinear models varied consistently across different choices of initial values of the fit. Also, the efforts of modeling the so-called volatility observed in financial time series led to the developments of the autoregressive conditional heteroscedastic model (ARCH) and its variants; for an account, see Tsay (2005) and references therein. Moreover, the introduction of long memory processes (Granger and Joyeux, 1980) were motivated by the need of describing the persistent correlation observed in many real phenomena. Lastly, recent development in nonparametric and semiparametric regression methods gave the possibility of modeling nonlinear phenomena with less assumptions on the data generating process; see Fan and Yao (2003), Gao (2007), and references therein.

The need for a different approach to time series analysis has been advocated by many authors, but prior to entertain the difficult task of nonlinear modeling it is sensible to make sure that a linear representation is not appropriate. This problem motivates the introduction of tests for assessing the presence of nonlinearity. In this chapter, we present a review on the literature of testing for nonlinearity in time series. The problem can be described as follows: *We are given a finite time series* $\mathbf{x} = (x_1, \ldots, x_n)$ *realization of a (strictly) stationary stochastic process* $\{X_t\}_{t\in\mathbb{Z}}$. *Now, on the basis of* $\mathbf{x}$ *we would like to assess with a certain confidence whether the process* $\{X_t\}_{t\in\mathbb{Z}}$ *that has generated the data is linear.*

The statement can be rewritten in terms of hypothesis testing:

$$\begin{cases} H_0 : \{X_t\}_{t\in\mathbb{Z}} \text{ is a linear process} \\ H_1 : \{X_t\}_{t\in\mathbb{Z}} \text{ is not a linear process} \end{cases} \tag{1}$$

In order to test this hypothesis, we need to define precisely the class of linear processes from a mathematical point of view. Also, we need to specify in some sense the meaning of H_1, that is, we need to describe the class of processes that cannot be represented in terms of H_0. Since there can be departures from linearity in many directions, testing the hypothesis (1) often becomes a test on a specific nonlinear feature. Such features have been found in phenomena from different disciplines and this has determined a fruitful cross-fertilization. For instance, some of the peculiar notions of nonlinear dynamics and chaos theory have motivated the introduction of new tools for time series analysis. In other situations, the nonlinearity can be seen as the inability of a linear model to describe the serial dependence observed in the data. Thus, the problem reduces to either a diagnostic test (often performed on the residuals of a linear model) or a specification test between models. On the basis of the above considerations, we have tried a classification of the various tests proposed in the vast literature that spreads across different disciplines. Of course, the borders between the various classes are blurred and a different schematization is possible.

The chapter is structured as follows. Section 2 is devoted to the definition of a linear process; we examine the practical implications of the mathematical representations and discuss the departures from linearity under the perspective of prediction. Section 3 presents the various tests classified into: tests based on the bispectrum and higher order moments (Section 3.1), diagnostic tests (Section 3.2), specification tests and Lagrange Multiplier tests (Section 3.3), nonparametric tests (Section 3.4), tests based on chaos theory (Section 3.5), tests based on surrogate data (Section 3.6). The last section presents a brief discussion and the conclusions.

2. Defining a linear process

A linear stationary process $\{X_t\}_{t\in\mathbb{Z}}$ is usually defined as

$$X_t = \sum_{j=1}^{\infty} \psi_j \varepsilon_{t-j} + \varepsilon_t, \tag{2}$$

where $\{\varepsilon_t\}$ is an *i.i.d.* process, $E[\varepsilon_t] = 0$, $Var[\varepsilon_t] = \sigma^2 < \infty$ and $\sum_{j=0}^{\infty} \psi_j^2 < \infty$. Hence a linear process admits a moving average (MA(∞)) representation. Also, if the MA transfer function $\Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j$ exists and has no zeros in $|z| \leq 1$ ($z \in \mathbb{C}$) then such processes admit an autoregressive (AR(∞)) representation:

$$X_t = \sum_{j=1}^{\infty} \phi_j X_{t-j} + \varepsilon_t, \tag{3}$$

where the coefficients $(\phi_j)_{j\in\mathbb{N}}$ are given by $1/\Psi(z) = 1 - \sum_{j=0}^{\infty} \phi_j z^j$.

Now, given these representations, can we hope to test efficiently the hypothesis (1)? The answer is not so clear cut as pointed out in Bickel and Bühlmann (1996, 1997). In fact, the authors define a topology over the set of stochastic processes by using different metrics and prove that both the set of MA processes and that of invertible AR processes

are not closed. Indeed, the closure of such two sets is quite large and comprises three kinds of processes: (i) the set of zero mean stationary Gaussian processes; (ii) the set of MA processes as defined in (2); (iii) a set of nonergodic processes that are Poisson sums of *i.i.d.* copies of a stationary process. Now, given any (even infinitely long) realization of a stationary process $(\xi_t)_{t\in\mathbb{Z}}$, define the process $\{X_t\}_{t\in\mathbb{Z}}$, a member of class (iii) defined above: $X_t = \sum_{j=1}^{N} \xi_{t;j}$, $(t \in \mathbb{Z})$ with $N \sim$ Poisson (1) and where $\xi_{t;j}$ are $i.i.d.$ copies of ξ_t for $j = 1, 2, \ldots$. Now, it can be shown that $\{X_t\}_{t\in\mathbb{Z}}$ is a member of the MA closure. Since $P[N = 1] = e^{-1} > 0.36$, we have that $P[X_t = \xi_t, \forall t] > 0.36$, almost surely. This leads to the following interesting fact:

FACT 1. Suppose we want to test the hypothesis H_0 that the observed series is a realization of a process of the class MA(∞) (2). Then, there is no test with asymptotic significance level $\alpha < 0.36$ that has limiting power one as $n \to \infty$.

In other words, even with infinite time series, it is impossible to distinguish perfectly between linear and nonlinear processes or, put it in another way, given a finite series, it is always possible to find a good description for it by means of a linear model, possibly of sufficiently high order. Incidentally, these results are at the basis of the sieve bootstrap methodology (Bühlmann, 1997). Of course, this does not mean that we should give up our endeavours. In fact, there exist subsets of the MA and AR classes that are closed and for which we can hope to build powerful tests. These subsets are sufficiently large as to include the linear models used in practice. For instance, it is often assumed that the data generating process under H_0 is a minimum phase finite order ARMA model of the kind:

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \tag{4}$$

where $\{\varepsilon_t\}$ is an *i.i.d.* process, $E[\varepsilon_t] = 0$, $Var[\varepsilon_t] = \sigma^2 < \infty$. By minimum phase it is meant that AR and MA polynomials associated to the above model have their zeros outside the closed unit circle in the complex plane. This implies that the stationary solution X_t is causal and invertible. Also, it is long known that, in some cases, one can use a linear model of high order for modeling nonlinear processes, so that what is the advantage of looking for nonlinear models instead? The key is *dimension reduction*. Indeed, nonlinear models are able to encompass complex features with few degrees of freedom. This, in turn, reflects on our capability of finding a parsimonious representation for a phenomenon. Moreover, some features cannot be adequately represented by linear models, no matter their order. As we will show in the following sections the notion of dimensionality of a process is at the basis of many tests for nonlinearity motivated by chaos theory.

Other fundamental aspects of linear processes are related to prediction, Gaussianity, and the notion of reversibility. A process is reversible if its joint probability structure is the same if we consider it with time reversed. It is clear that Gaussian stationary processes are reversible but many non-Gaussian processes are not. Much of the time series literature is based on Gaussian models and methods motivated by their paradigm. However, even non-Gaussian linear stationary processes present richer and more complex features than those typical of linear Gaussian processes. For this reason it seems reasonable to test, among other things, whether we are dealing with either linear Gaussian

processes or linear non-Gaussian processes. Assume that $\{X_t\}_{t\in\mathbb{Z}}$ is a minimum phase stationary ARMA process as in Eq. (4). Consider the prediction problem in which one approximates X_t by a function of the past $(X_{t-1}, X_{t-2}, \dots)$. Then, the best predictor of X_{t+m} in the mean square sense is given by the conditional expectation

$$E[X_{t+m}|X_s, s \leq t] \qquad m \in \mathbb{N} \tag{5}$$

Now, it is known that the conditional expectation of Eq. (5) is a linear function of $\{X_s, s < t\}$ when $\{X_t\}_{t\in\mathbb{Z}}$ is Gaussian. Moreover, in the minimum phase case such predictor has the same linear form as in the Gaussian case, no matter the distribution of ε_t. In the general case, however, if the distribution of ε_t is not Gaussian without additional constraints on the process, such conditional expectation is a nonlinear function of lagged variables, see for instance Tong (1990, p. 13) and references therein. For further details and discussions on linear non-Gaussian processes, see also Rosenblatt (2000). The author also shows that given a stationary AR(1) process: (i) the best mean square predictor forward in time is linear; (ii) the best mean square predictor backward in time is linear if and only if ε_t follows a Gaussian distribution. Hence, the notion of reversibility seems intimately related to linear Gaussian processes. Again, these findings prompted studies on tests for reversibility as we will show in Section 3.5. For a systematic treatment of the relationship between linear representations and prediction, see Chapter 1 of Hannan and Deistler (1988), Chapter 5.5 of Pourahmadi (2001), and also Chapter 5 of Brockwell and Davis (1991).

2.1. *What is a nonlinear process?*

The question comes immediately to our minds: can we define mathematically a nonlinear process in the same way we have defined a linear one? The answer is negative. As pointed out above there is an intrinsic asymmetry between the two realms. Since there can be departures from linearity in various directions, we can only define a nonlinear phenomenon through those features that cannot be exhibited by linear processes and that have been observed in various disciplines. Among such features we can mention asymmetry, regime effects, presence of limit cycles, irreversibility, nonuniform predictability, noise amplification/suppression, phase synchronization, noise-induced phenomena, and sensitivity to initial conditions. In most cases, the need for describing these behaviors led to the introduction of new tests and models that paved the way for major advances in our understanding of the nonlinear world. Extended accounts from a time series analysis perspective can be found in the studies done by Chan and Tong (2001), Tong (1990), Fan and Yao (2003), and Gao (2007), while from the point of view of nonlinear dynamical system theory one can refer to the works done by Kantz and Schreiber (2004), Abarbanel (1996), Broer and Takens (2011), Galka (2000), and Diks (1999).

Now, we will pursue a bit further the discussion on the prediction problem in the case of nonlinear processes and will show that the notions of sensitivity to initial conditions and nonuniform noise amplification/predictability emerge naturally. As shown in Yao and Tong (1994a,b), Chan and Tong (2001), and Fan and Yao (2003) the prediction problem presented above changes if the process is nonlinear. Assume we want to

predict X_{t+m} based on the past d observations: $\mathbf{X}_t = (X_t, \ldots, X_{t-d+1})$. It is immediate to see that the least square predictor for X_{t+m} is

$$f_{t,m}(\mathbf{x}) = E[X_{t+m}|\mathbf{X}_t = \mathbf{x}], \tag{6}$$

furthermore, the mean square prediction error is given by

$$E[(X_{t+m} - f_{t,m}(\mathbf{x}))^2] = E[\sigma^2_{t,m}(\mathbf{x})], \tag{7}$$

where $\sigma^2_{t,m}(\mathbf{x}) = Var[X_{t+m}|\mathbf{X}_t = \mathbf{x}]$. Such measure monitors the performance of the prediction and is constant if the process is a linear AR(p). In the general case, however, the goodness of the prediction depends on the initial state of the system. Furthermore, Yao and Tong (1994b) discussed the notion of initial value sensitivity in a stochastic environment and showed that, in case of a nonlinear system, a small uncertainty on the initial condition can have an impact on the prediction error. Suppose we try to predict X_{t+m} by $f_{t,m}(\mathbf{x})$ where $X_t = \mathbf{x}$. Now, assume we do not know $\mathbf{x}$ exactly but we are subject to a small error $\boldsymbol{\delta}$ such that $X_t = \mathbf{x} + \boldsymbol{\delta}$. This assumption is natural since, in practice, we never know exactly the state of the system. Now, Yao and Tong (1994b) proved the following decomposition theorem:

$$E[(X_{t+m} - f_{t,m}(\mathbf{x}))^2|\mathbf{X}_t = \mathbf{x} + \boldsymbol{\delta}] = \sigma^2_{t,m}(\mathbf{x} + \boldsymbol{\delta}) + \{\boldsymbol{\delta}' \dot{f}_{t,m}(\mathbf{x})\}^2 + o(||\boldsymbol{\delta}||^2). \tag{8}$$

This result shows that the prediction performance depends on (i) the conditional variance $\sigma^2_{t,m}(\mathbf{x})$ that measures the amount of randomness, and (ii) the uncertainty in the initial condition $\boldsymbol{\delta}$ through $\dot{f}_{t,m}(\mathbf{x})$, the gradient vector of $f_{t,m}(\mathbf{x})$. This nonuniform noise amplification is related to initial value sensitivity and is a peculiar feature of nonlinear processes. This results has an important consequence on multistep prediction. In the linear case, $\dot{f}_{t,m}(\mathbf{x})$ is constant and the remainder of the right hand side of Eq. (8) is zero. Also, $\sigma^2_{t,m}(\mathbf{x})$ does not depend on $\mathbf{x}$ and the mean square predictive error increases monotonically with m. This is not always the case with nonlinear prediction so that we might be able to predict $m + 1$ steps ahead better than m steps ahead. This phenomenon has been shown empirically in the study by Subba Rao and Gabr (1984), where the authors computed the prediction error variance for different nonlinear models of the time series of the Canadian Lynx and of the Sunspot index. For further discussions, see Fan and Yao (2003) and Chan and Tong (2001).

3. Testing for nonlinearity

Needless to say, the plethora of tests for nonlinearity, or nonlinear serial dependence, in time series is vast. In this section we will try to discuss the various proposals and make a tentative classification. As also pointed out in the works of Barnett et al. (1997) many of the available tests have different hypotheses, both null and alternative, so that there is little point in comparing them. Rather, some of them can be used jointly as the hypotheses tested are completely different.

3.1. Tests based on the bispectrum and higher order moments

The first account on a test for nonlinearity in the time series literature is probably that of Subba Rao and Gabr (1980). The proposal is based upon the properties of the bispectrum. In practice, such approach tries to assess the linearity and the Gaussianity of a series by looking at third-order moments. The procedure can also be interpreted as testing for the significance of the coefficients associated to the linear terms of the Wiener expansion of the solution of the process (a stochastic version of the Volterra series representation). The tests can be applied both to the original series and to the residuals of a fitted model. Assume that $\{X_t\}_{t\in\mathbb{Z}}$ is a sixth-order stationary process with zero mean, covariance function $\gamma_k = E[X_t X_{t+k}]$, $k \in \mathbb{Z}$ and spectral density function $f(\omega)$, $|\omega| \leq \pi$. The third-order cumulants can be written as $\gamma_{m,n} = E[X_t, X_{t+m}, X_{t+n}]$. Now, $\gamma_{m,n}$ is called the bicovariance function, whereas the bispectrum function $f(\omega_1, \omega_2)$ is defined as the double Fourier transform of the bicovariance:

$$f(\omega_1, \omega_2) = \frac{1}{(2\pi)^2} \sum_{m=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} \gamma_{m,n} e^{-i2\pi(\omega_1 m + \omega_2 n)}, \quad -\pi \leq \omega_1, \omega_2 \leq \pi.$$

In the same way $f(\omega)$ is a Fourier decomposition of $E[X_t^2]$, $f(\omega_1, \omega_2)$ is a frequency decomposition of the third moment $E[X_t^3]$ of the process $\{X_t\}_{t\in\mathbb{Z}}$. Now, in case of a linear process that admits a MA(∞) representation as that of Eq. (2), we have

$$X_{ij} = \frac{|f(\omega_i, \omega_j)|^2}{f(\omega_i) f(\omega_j) f(\omega_i + \omega_j)} = \frac{(E[\varepsilon_t^3])^2}{2\pi E[\varepsilon_t^2]} \quad \forall i, j. \tag{9}$$

The approach of Subba Rao and Gabr (1980) is based on the following two facts: (i) if $\{X_t\}_{t\in\mathbb{Z}}$ is a linear Gaussian process then $f(\omega_1, \omega_2) = X_{ij} = 0$, for all i, j and $E[\varepsilon_t^3] = 0$; (ii) if $\{X_t\}_{t\in\mathbb{Z}}$ is a linear non-Gaussian processes X_{ij} is constant for all the frequencies i, j. These facts lead to the definition of two statistical tests of Gaussianity and linearity that are mainly based on the following statistic: $S = 2\sum_{m,n} |\hat{X}_{m,n}|^2$, where $\hat{X}_{m,n}$ is an estimator of $X_{m,n}$ of Eq. (9). The asymptotic distribution of such statistic is a central χ^2 under the hypothesis of Gaussianity and a noncentral χ^2 distribution under the hypothesis of linearity. Now, for the latter case, Subba Rao and Gabr (1980) propose an F test for constant means of $2|X_{m,n}|^2$. As an alternative, Hinich (1982) proposes using the sample interquartile range of $2|X_{m,n}|^2$ (see also Brockett et al. (1988)). This latter proposal appears to have better properties than the original test in a number of situations; still, it is bounded to the choice of a smoothing parameter. See also Berg et al. (2010, 2012) for bootstrap versions of the test by Hinich (sieve and subsampling, respectively) and for a review of other versions of the bispectral tests. For further discussion on the properties of such test see also Ashley et al. (1986). Another test based on the bispectrum is proposed in the works done by Terdik and Math (1998). Here, the authors test the null hypothesis that the best predictor is linear against the alternative that such predictor is quadratic. Finally, Rusticelli et al. (2009) propose bispectral tests for linearity based on a maximization procedure as to eliminate the arbitrariness concerning the smoothing parameter; the test statistic proposed is derived by maximizing the classical Hinich statistic over a range of feasible values for

the smoothing parameter. Such tests appear to have more power than classical Hinich's tests, in particular with respect to NLMA (nonlinear moving average), GARCH and deterministic chaotic processes.

The time-domain counterpart of the bispectrum, namely, the bicovariance, can be used to build tests of nonlinear dependence that are similar in spirit to those based upon the bispectrum. See Barnett and Wolff (2005), Brooks and Hinich (2001), and references therein for an account. Moreover, in the study by Subba Rao and Wong (1998) multivariate measures of skewness and kurtosis are used for deriving tests for linearity and Gaussianity in vector time series. Also, in the works of Subba Rao (1992), the authors study possible applications of the bispectrum to non-Gaussian and chaotic time series. In particular, they show how the estimated higher order spectra could be used to distinguish between nonlinear deterministic stable systems and nonlinear deterministic chaotic systems. The authors observed that, compared to a stable signal, the energy of the estimated bispectrum of a chaotic signal is distributed over a broader range of frequencies.

It is worth noting that tests based on third-order moments concentrate on symmetry and cannot detect a nonlinearity that depends upon moments higher than the third. In theory, in order to rule out the presence of nonlinearities, all the cumulants should be tested. This is the motivation at the basis of tests that involve the complete joint distribution and that will be presented in the following.

3.2. Diagnostic tests

Many of the tests proposed in literature can be thought as "diagnostic tests" since they are applied on the residuals of a linear model such as that of Eq. (4). The idea is to regress the residual of a linear model on specific functions of X_t, chosen to capture essential features of possible nonlinearities; the null hypothesis is rejected if these functions of X_t are significantly correlated with the residuals. In this respect, in some instances, testing for neglected nonlinearity is tantamount to testing for a null of independence versus the alternative of serial dependence. The literature on tests of serial dependence (against the null of independence) is vast and its review is outside the scope of this chapter; rather we concentrate on specific proposals that have been used for diagnostic checking of time series models. Some of the classical tests are described in Chapters 5.3.2–5.3.3 of Tong (1990) and references therein. These include the Ljung-Box test, the Mcleod and Li test, the Keenan test, and the Tsay test. Notice that in many of such tests the alternative hypothesis is not clearly stated and Tong classifies them as portmanteau tests. For a detailed account on both diagnostic and goodness-of-fit tests for time series, see Li (2004).

Among the proposals we mention the BDS test (Brock et al., 1986), motivated by chaos theory and based upon the asymptotic distribution of the sample correlation integral. The correlation integral is defined as

$$C_d(\epsilon) = P\left[\prod_{j=0}^{d-1} \mathbb{I}(|X_{t-j} - X_{s-j}| < \epsilon)\right], \tag{10}$$

whereas its sample version is

$$\hat{C}_d(\epsilon) = \frac{2}{n(n-1)} \sum_{t=d+1}^{n} \sum_{s=d}^{t-1} \prod_{j=0}^{d-1} \mathbb{I}(|X_{t-j} - X_{s-j}| < \epsilon), \tag{11}$$

where $\mathbb{I}$ is the indicator function, d is the embedding dimension (or order of the state vector), and $\epsilon \in \mathbb{R}^+$ is the radius of the hypersphere. The statistic $\hat{C}_d(\epsilon)$ measures the proportion of pairs of phase space points that lie within a radius ϵ.

The test has the following form:

$$\text{BDS}(d, \epsilon) = \sqrt{n} \left[\hat{C}_d(\epsilon) - \hat{C}_1(\epsilon)^d \right] / \hat{V}_d^{1/2} \tag{12}$$

with $\hat{V}_d$ being an estimator for the asymptotic variance. Under the null of independence we have $C_d(\epsilon) = C_1(\epsilon)^d$. Basically, the test it is aimed at detecting departures from independence within a range of lags/embedding dimensions specified by the experimenter. Notice that the results of this test depend sensibly from the choice of the embedding dimension d and of the radius of the sphere ϵ. Also, the "nuisance parameter free" condition claimed by the authors holds only under conditional mean models but not under ARCH-type models. Moreover, serial independence implies $C_d(\epsilon) = C_1(\epsilon)^d$ but the converse is not necessarily true.

Recent proposals involve tests on the joint pairwise distribution of a (strictly) stationary stochastic process $\{e_t\}$, which is supposed to represent the process behind the standardized residuals of a fitted model. In particular, the idea of Hong (1999), Hong and Lee (2003) is to consider the spectrum of the transformed series $\{e^{iue_t}\}$, where $u \in \mathbb{R}$. First, define the covariance function at lag j as $\sigma_j(u, v) = Cov(e^{iue_t}, e^{ive_{t-j}})$, with $j \in \mathbb{Z}$ and $i = \sqrt{-1}$. Clearly, $\sigma_j(u, v) = \varphi_j(u, v) - \varphi_j(u)\varphi_j(v)$ where $\varphi_j(u, v)$ and $\varphi_j(u)$ are the joint and marginal characteristic functions of (e_j, e_{t-j}). Hence, $\sigma_j(u, v) = 0$ if and only if e_j and e_{t-j} are independent. Under mild conditions on $\{e_t\}$ we have that the Fourier transform of $\sigma_j(u, v)$ exists:

$$f(\omega, u, v) = \frac{1}{2\pi} \sum_{j=-\infty}^{+\infty} \sigma_j(u, v) e^{-ij\omega}, \quad -\pi \leq \omega \leq \pi.$$

$f(\omega, u, v)$ can capture any kind of pairwise dependence across various lags in e_t. For instance, the negative partial derivative of $f(\omega, u, v)$ with respect to (u, v) at $(0, 0)$ yields the conventional spectral density. For such reasons the authors denote $f(\omega, u, v)$ as a generalized spectral density. Now, it can be proven that under the null of a $i.i.d.$ process the generalized spectral density is flat: $f_0(\omega, u, v) = \frac{1}{2\pi}\sigma_0(u, v)$. On this basis, Hong and Lee (2003) derive a diagnostic test for neglected nonlinearity by using a measure of L_2 divergence between the sample version of $f(\omega, u, v)$ and $f_0(\omega, u, v)$. Interestingly, no moment condition on $\{e_t\}$ is required and this is a desirable property in many situations (i.e., high-frequency financial time series). Moreover, when applied to the standardized residuals of a wide class of models, the test has an asymptotic distribution that is free from nuisance parameters. In other words, the limit distribution

under the null of the test statistic does not depend on the estimators of the parameters of the fitted models, provided that such estimators are $n^{1/2}-$consistent.

Another class of diagnostic tests is based on nonparametric entropy measures of dependence motivated by information theory. Ideally, such measures can be used as nonlinear autocorrelation functions and overcome the known limitations of the linear correlation coefficient. For a review and a discussion on the relevant asymptotic theory, see Tjøstheim (1996) and Hong and White (2005). Robinson (1991) uses the Kullback–Leibler information divergence as a basis for one-sided testing of nested hypotheses. Kernel density estimation is used to derive the test statistic and in order to obtain a normal null limiting distribution, a form of weighting is introduced. In the study of Granger and Lin (1994) the *mutual information* is used, whereas in the study of Granger et al. (2004) a discussion on the axiomatic properties of an ideal measure of dependence is put forward. The discussion leads to adopting the metric entropy measure S_ρ, a normalized version of the Bhattacharya–Hellinger–Matusita distance, defined as follows:

$$S_\rho(k) = \frac{1}{2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left[\sqrt{f_{(X_t,X_{t+k})}(x_1,x_2)} - \sqrt{f_{X_t}(x_1)f_{X_{t+k}}(x_2)}\right]^2 dx_1 dx_2, \quad (13)$$

where $f_{X_t}(\cdot)$ and $f_{(X_t,X_{t+k})}(\cdot,\cdot)$ denote the probability density function of X_t and of the vector (X_t, X_{t+k}), respectively. The measure is a symmetrized general "relative" entropy, which includes as a special case nonmetric relative entropies such as the Kullback–Leibler divergence. $S_\rho(k)$ satisfies many desirable properties, in particular, (i) it is a metric; (ii) it is normalized and takes the value 0 if X_t and X_{t+k} are independent and 1 if there is a measurable exact relationship between continuous variables; (iii) it reduces to a function of the linear correlation coefficient in the case of Gaussian variables. Notably, we have that $\{X_t\}_{t\in\mathbb{Z}}$ is an independent process if and only if $S_\rho(k) = 0$ for all $k \neq 0$. As in the works done by Granger and Lin (1994), a kernel density implementation of S_ρ has been adopted, leading to the following estimator:

$$\hat{S}_\rho(k) = \frac{1}{2}\int\int\left[\sqrt{\hat{f}_{(X_t,X_{t+k})}(x_1,x_2)} - \sqrt{\hat{f}_{X_t}(x_1)\hat{f}_{X_{t+k}}(x_2)}\right]^2 w(x_1,x_2)dx_1dx_2, \quad (14)$$

where the densities are estimated through kernel regression and $w(x_1, x_2)$ is a weight function that is needed in order to derive the asymptotic distribution of the estimator. In the study by Maasoumi and Racine (2009), the measure is used to build a test for asymmetry both for discrete and continuous processes, whereas, in the study by Granger et al. (2004), a permutation framework is used to test the null of independence; the superiority of tests based on S_ρ as compared to both the Ljung-Box and the BDS tests is shown. Giannerini et al. (2007a) extend the results of Granger et al. (2004) and use the entropy measure to build tests for nonlinearity based on different resampling schemes. Lastly, Fernandes and Néri (2010) discuss various entropy measures to derive tests for independence between stochastic processes.

Finally, we mention a diagnostic test for linearity proposed by An and Cheng (1991). The idea is to derive a Kolmogorov–Smirnov type test for linearity from the residuals of a fitted AR model. Under a similar setup, Lobato (2003) defines Cramér–Von Mises and Kolmogorov–Smirnov type statistics for deciding whether the conditional mean of X_t is a linear autoregression of finite order. The device proposed makes use of a sequence of alternatives that tends to the null hypothesis at a rate $n^{-1/2}$. The asymptotic distribution of the test statistic is found by means of bootstrap methods.

3.3. *Specification tests and lagrange multiplier tests*

This section reviews briefly those tests that aim at assessing the null of linearity against alternatives of specific nonlinear models. These tests are usually more mathematically involved, but with respect to the specific alternatives they also give higher power than pure significance tests. The task can be accomplished by means of either parametric or nonparametric methods. In the first case, typically, the following model is hypothesized:

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + f(\boldsymbol{\beta}, X_{t-1} + \cdots + X_{t-p}, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}) + \varepsilon_t. \tag{15}$$

The model consists of a linear (ARMA) part and a nonlinear part f that depends upon an unknown parameter vector $\boldsymbol{\beta}$. Clearly, testing for linearity amounts to assessing whether $\boldsymbol{\beta} = 0$. This can be done by using a Lagrange multiplier (LM) approach as put forward in the study by Luukkonen et al. (1988) and Saikkonen and Luukkonen (1988). The general framework of the LM test can be used to test a linear model against different parametric forms for f. These include, for instance, ARCH and GARCH models (Engle, 1982), Bilinear models, SETAR models, EXPAR models, see Tong (1990), Chapter 5.3.5 and references therein. More recently, a neural network version of the LM test has been proposed (Lee et al., 1993). The idea is to model the nonlinear function f of Eq. (15) by means of a single hidden layer network of the kind

$$f(\cdot) = \sum_{j=1}^{k} \beta_{0j} \left\{ \psi(\mathbf{w}_j' \mathbf{X}_t) - \frac{1}{2} \right\} + \varepsilon_t, \tag{16}$$

where $\varepsilon_t \sim WN(0, \sigma^2)$, $\mathbf{w}_j = (w_{0j}, w_{1j}, \ldots, w_{hj})'$, and $\mathbf{X}_t = (1, X_{t-1}, \ldots, X_{t-h})'$. Hence, the null hypothesis of the neural network test against neglected nonlinearity is H_0: $\beta_{01} = \cdots = \beta_{0k} = 0$. Notice that, in general, the functional form of ψ is unknown and the nonlinear model is only identified under the alternative; the problem has been faced in the works of Teräsvirta et al. (1993), where they focus on the special case where $k = 1$, that is, the net has 1 unit and $\mathbf{w} = (w_0, w_1, \ldots, w_h)'$. The authors use a Taylor expansion of $\psi(\mathbf{w}'\mathbf{X}_t)$ around $\mathbf{w} = 0$. This leads to testing the following hypothesis H_0: $w_1 = \cdots = w_h = 0$ that, in turn is equivalent to

$$H_0: \delta_{ij} = 0; \delta_{ijl} = 0; \quad i = 1, \ldots, h;\ j = i, \ldots, h;\ l = j, \ldots, h.$$

where

$$f(\cdot) = \sum_{i=1}^{k}\sum_{j=1}^{k}\delta_{ij}X_{t-i}X_{t-j} + \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{l=1}^{k}\delta_{ijl}X_{t-i}X_{t-j}X_{t-l} + \varepsilon_t.$$

This version of the neural network test based on the Volterra series expansion is free of unidentified nuisance parameters under the null and has better power properties than that of the works done by Lee et al. (1993). For another battery of LM tests that avoids the problem of unidentified nuisance parameters under the null, see Dahl and Gonzlez-Rivera (2003). The proposal is based on random fields theory.

A great deal of attention has been placed on testing for linearity against threshold models. Besides the above proposals based on LM tests the relevant theory and a discussion of such tests is presented in detail in the study by Tong (1990) and Li (2004), and references therein.

3.4. Nonparametric tests

In this section, we present some nonparametric tests for nonlinearity that cannot be classified as diagnostic tests since they are applied directly on the original series. Also, such approaches allow to identify the lags at which a nonlinear effect is expected so that they might provide useful information for the subsequent specification of a nonlinear model. The first two related proposals are those of Hjellvik and Tjøstheim (1995) and Hjellvik et al. (1998). The main idea of Hjellvik and Tjøstheim (1995) relies on comparing the linear and the nonlinear least square predictors for X_t given X_{t-k}. Given a zero mean stationary process with finite fourth order moments and autocorrelation function at lag k ρ_k the procedure is based on $M_k(x) = E[X_t|X_{t-k} = x]$ and $V_k(x) = V[X_t|X_{t-k} = x]$. The test statistics are:

$$L(M_k) = E\left[\{M_k(X_{t-k}) - \rho_k X_{t-k}\}^2\right] \tag{17}$$

$$L(V_k) = E\left[\{V_k(X_{t-k}) - (1-\rho_k^2)V(X_{t-k})\}^2\right] \tag{18}$$

Now, $M_{k(x)}$ and $V_{k(x)}$ are estimated through kernel (local constant) regression so that, in practice, the tests compare nonparametric and parametric estimators. Clearly, under the null of a linear Gaussian process $L(M_k) = L(V_k) = 0$ for all k. Since asymptotic results require very long series to give meaningful results, the authors propose a sieve bootstrap approach for deriving the rejection regions.

In the study by Hjellvik et al. (1998) similar arguments are put forward, this time exploiting local polynomial estimators of $M_k(x)$ and its derivatives. Moreover, the authors do not assume that the noise process is $i.i.d.$ so that the conditional heteroscedastic case is accounted for. The null hypothesis tested is

$$\begin{cases} H_0 : M_k(x) \text{ is linear } \forall k = 1, \ldots, l \\ H_1 : M_k(x) \text{ is nonlinear for at least one } k \end{cases} \tag{19}$$

where l is the maximum lag for which the test is performed. Notice that if the data generating process is linear then

$$M_k(x) = \rho_k x; \quad M_k'(x) = \rho_k; \quad M_k''(x) = 0 \quad \forall x$$

furthermore:

$$V_k(x) = c; \quad V_k'(x) = 0 \quad \forall x,$$

where c is a constant. These facts motivate the following test statistics:

$$\begin{aligned}
L(M_k) &= E\left[\{M_k(X_{t-k}) - \rho_k X_{t-k}\}^2\right] \\
L(M_k') &= E\left[\{M_k'(X_{t-k}) - \rho_k\}^2\right] \\
L(M_k'') &= E\left[\{M_k''(X_{t-k})\}^2\right] \\
L(V_k) &= E\left[\{V_k(\hat{e}_{t-k}) - 1)V(\hat{e}_{t-k})\}^2\right] \\
L(V_k') &= E\left[\{V_k'(\hat{e}_{t-k})\}^2\right]
\end{aligned}$$

where $\hat{e}_t$ are the residuals from fitting a linear model. As in the previous case, the distribution of the statistics under the null hypothesis is derived by means of the sieve bootstrap. Notice that in the above mentioned cases the alternative hypothesis is highly composite so that it is not possible to formulate theoretical assumptions on the power of the tests.

A third nonparametric test for nonlinearity is based on the entropy metric S_ρ of Eq. (13), as proposed in the work of Giannerini et al. (2011). The idea is somehow similar to that of Hjellvik and Tjøstheim (1995), Hjellvik et al. (1998) in that the test statistic is based on the divergence between a parametric (linear) and a nonparametric (unrestricted) estimator of the same quantity. Here, the hypothesis tested is that of a zero mean linear (Gaussian) process. The authors prove that under H_0 $S_\rho(k)$ reduces to a smooth bounded function of the autocorrelation coefficient at lag k as follows:

$$S_\rho(k) = 1 - \frac{2\left(1-\rho_k^2\right)^{1/4}}{\sqrt{4-\rho_k^2}}. \tag{20}$$

Hence, the following test statistic is proposed:

$$\hat{T}_k = \left[\hat{S}_k^u - \hat{S}_k^p\right]^2. \tag{21}$$

$\hat{T}_k$ is the squared divergence between the unrestricted nonparametric estimator $\hat{S}_k^u$ (see Eq. 14) and the parametric estimator of $S_\rho(k)$ under H_0 based on Eq. (20) and denoted by $\hat{S}_k^p$. The authors prove that under H_0:

1. $\hat{T}_k \xrightarrow{p} 0$ (22)
2. $\frac{n\hat{T}_k}{\sigma_a^2} \xrightarrow{d} \chi_1^2,$ (23)

where σ_a^2 is the asymptotic variance of $\sqrt{\hat{T}}_k$. As in previous cases, since asymptotic theory is not applicable in practice, the authors derive the distribution of $\hat{T}_k$ under the null in finite samples by means of both the sieve bootstrap and surrogate data. This latter approach will be presented in the next sections.

3.5. Tests based on chaos theory

One of the peculiar features of chaotic systems is the well-known sensitive dependence on initial conditions, i.e., infinitesimal perturbations on the initial state of the system are exponentially amplified. This notion, introduced in the field of nonlinear dynamical system, has gained attention in the Statistics community since it provides new tools and concepts for characterizing nonlinear time series. In the following, we will review the literature on testing for the presence of initial value sensitivity in a time series context.

Assume that the data generating process is a stochastic difference equation of the kind

$$\mathbf{X}_{t+1} = F(\mathbf{X}_t) + \mathbf{e}_{t+1}, \quad t \in \mathbb{Z}^+, \tag{24}$$

where $\mathbf{X}_t = (X_t, \ldots, X_{t-d+1})$, $F\colon \mathbb{R}^d \to \mathbb{R}^d$, and $_t$ is a $i.i.d.$ d-dimensional random process. Notice that the stochastic component is a part of the process and interacts with the deterministic skeleton. Also, when the noise term is negligible we treat the system as deterministic. Now, it is possible to show that Eq. (24) implies:

$$X_{t+1} = f(\mathbf{X}_t) + \varepsilon_{t+1}, \quad t \in \mathbb{Z}^+, \tag{25}$$

where ε_t is a noise process with $E[\varepsilon_t | X_{t-k}] = 0$, $\sigma^2 = Var(\varepsilon_t) = Var(\varepsilon_t | X_{t-k})$ with $k > 0$.

One of the measures of initial value sensitivity for deterministic systems is the maximal Lyapunov characteristic exponent (MLCE). The MLCE measures the average rate of divergence of trajectories with nearby initial conditions. It is an important measure of stability and one of the indicators of the presence of chaos. In fact, a positive MLCE is a necessary condition for the presence of chaos. Denote with $\mathbf{X}_0$ and $\mathbf{X}'_0$ two close initial conditions in the phase space and with $\mathbf{X}_n$ and $\mathbf{X}'_n$ their value after n iterations, respectively. Then, the MLCE can be defined as

$$\lambda = \lim_{n\to\infty} \lim_{\delta\to 0} \frac{1}{n} \ln \left(\frac{\left\| \mathbf{X}_n - \mathbf{X}'_n \right\|}{\left\| \mathbf{X}_0 - \mathbf{X}'_0 \right\|} \right) \tag{26}$$

where $\|\cdot\|$ is an appropriate norm and $\delta = \left\| \mathbf{X}_0 - \mathbf{X}'_0 \right\|$ is the perturbation in the initial condition. This definition holds with probability one for almost all initial conditions. Notice that if we denote with $F^{(n)}(\mathbf{x}_0)$ the n-th iteration of the system we obtain

$$\begin{aligned} \mathbf{X}_n - \mathbf{X}'_n &= F^{(n)}(\mathbf{X}_0) - F^{(n)}(\mathbf{X}'_0) \\ &\approx DF^{(n)}(\mathbf{X}_0)(\mathbf{X}_0 - \mathbf{X}'_0). \end{aligned} \tag{27}$$

Hence, the growth of an infinitesimal perturbation in the initial condition is governed by the Jacobian matrix of partial derivatives of the map F. If in Eq. (26) we do not take the limit $n \to \infty$ but we consider only a finite number of steps ahead k, we obtain the so called k-step ahead Local Lyapunov exponents (LLE), which can be used for characterizing the predictability in different regions of the state space. Estimation of Lyapunov exponents both for deterministic and stochastic systems are discussed in the works done by Giannerini and Rosa (2004). Notice that estimation of the map F is required. Formal tests for chaos where $H_0 : \lambda = 0$ against $H_1 : \lambda > 0$ can be found in the works of Shintani and Linton (2004) (neural network approach), Whang and Linton (1999) and Park and Whang (2012) (local polynomial regression approach). For a different approach valid for time continuous systems and based on spline interpolation, see Giannerini and Rosa (2001) and Giannerini et al. (2007b).

The problem of assessing sensitivity to initial conditions in the stochastic case can been faced by assuming that the state of the system is a random variable. Hence, instead of speaking of average rate of divergence of nearby starting trajectories (the MLCE) one can define divergence and initial value sensitivity of the conditional distribution of the system. This is the approach entertained by Yao and Tong (1994a,b) where measures of sensitivity of the conditional mean and conditional quantiles are derived. In the first case, these reduce to the classical MLCE in the limit of vanishing noise and are based on the gradient vector $\dot{f}_{t,m}(\mathbf{x})$ defined in Eq. (8). Measures based on the conditional density are introduced in the study by Fan et al. (1996). Notice that formal tests test based on such measures have not been proposed. Still, these have been used successfully for characterizing nonlinear time series in the works done by Chan and Tong (2001) and Giannerini and Rosa (2004).

Finally, it is worth of mention a test for reversibility motivated by chaos theory and presented in the study by Diks (1999). The author also discusses the estimation of invariant quantities for noisy dynamical systems, in particular he focuses on the correlation integral presented in Section 3.2.

3.6. Tests based on surrogate data

The method of surrogate data was introduced in the field of nonlinear dynamics and can be seen as a resampling approach for building tests for nonlinearity in time series. The work of Theiler et al. (1992) is usually indicated as the seminal paper on the subject. The main idea at the basis of the method can be summarized as follows: (i) a null hypothesis regarding the process that has generated the observed series (DGP) is formulated; for instance, H_0: the DGP is linear and Gaussian, (ii) a set of B resampled series, called *surrogate series*, consistent with H_0, are obtained through Monte Carlo methods, (iii) a suitable test statistic, known to have discriminatory power against H_0, is computed on the surrogates obtaining the distribution of the test statistic under H_0, (iv) the significance level of the test is derived by comparing the value of the test statistic computed both on the original series and on the surrogate distribution. Notice that the basic principle behind surrogate data tests is closely related to the bootstrap principle.

In the study by Theiler et al. (1992) and Theiler and Prichard (1996), a null hypothesis of linearity is tested by generating surrogates having the same periodogram and the same marginal distribution as the original series. In brief, surrogate series

$\mathbf{y} = (y_1, \ldots, y_n)^T$ are derived by randomizing the phases of the Fourier transform of the original series $\mathbf{x}$ as to obtain

$$y_t = \bar{x} + \sqrt{\frac{2\pi}{n}} \sum_{j=1}^{m} 2\sqrt{I(\mathbf{x}, \omega_j)} \cos(\omega_t j + \theta_j), \tag{28}$$

where $\bar{x}$ is the sample mean, $\omega_j = \frac{2\pi_j}{n}$, $j = 1, \ldots, n$ are the angular frequencies, $I(\mathbf{x}, \omega)$ is the sample periodogram and $\theta_1, \ldots, \theta_m$, $(m = (n-1)/2)$ are i.i.d. phases $U[0, 2\pi]$. For further details on the derivation of Eq. (28) see Chan (1997). The surrogate series will have the same sample mean and periodogram of the original series. The rationale behind the method is that, given the sample mean the periodogram values and the phases, one can always recover the original signal. Hence, by randomizing the phases one creates a signal that preserves the original mean and periodogram but loses all the remaining information.

The null hypotheses tested in this context are: (i): $\{X_t\}_{t\in\mathbb{Z}}$ is a linear process; (ii): $\{X_t\}_{t\in\mathbb{Z}}$ is nonlinear monotone transform of a linear process. This latter hypothesis is tested by imposing a further adjustment to phase randomized surrogates.

Since its first appearance, the method of surrogate data gained quite a lot of popularity among applied scientists and several extensions have been proposed; in general, these concern the introduction of either new test statistics or *ad hoc* algorithms for generating surrogates for testing specific (not necessarily linear) hypotheses. For instance, Small and Judd (1998), Small et al. (2001) and Small (2005) propose a class of statistics based on the correlation integral and prove their pivotalness with respect to the above hypotheses. Then, they apply the method in order to study infant sleep apnea, ECG dynamics and human vocalization patterns. In the works done by Galka (2000), the method of surrogate data is applied to EEG time series, and in the work of Dolan et al. (1999) a surrogate approach for finding unstable periodic orbits is proposed. Finally, in the study by Kugiumtzis (2002, 2008) an alternative surrogate generation scheme for testing hypothesis (ii) above is proposed and its performance assessed on both simulated data and EEG series.

Despite the great interest arisen, there are open theoretical issues that have been solved only partially or have not been considered at all. The first problem concerns the performance of the method. As several authors point out (see Schreiber and Schmitz (2000) and references therein), the phase randomization device usually leads to high false positive rates. The problem has been discussed in several instances, see, e.g., Kugiumtzis (2001), Galka (2000), Schreiber and Schmitz (1996), and Theiler and Prichard (1997). A proposal that partially overcomes this problems is that of Schreiber (1998), also discussed in the work done by Giannerini et al. (2011). In practice, surrogate generation is seen as a constrained stochastic optimization problem solved through simulated annealing. For an interesting review on the topic, see Schreiber and Schmitz (2000) and references therein.

A second problem regards the validity of the method. The first rigorous results in this direction are due to Chan (1997), which shows that the phase randomization method described above is (i) exactly valid under the null hypothesis that the DGP is a stationary Gaussian circular process; (ii) asymptotically valid for the null hypothesis that the DGP is a stationary Gaussian process with fast-decaying autocorrelations. By

valid, it is meant that tests have a Neyman structure (see also Chan and Tong (2001), Chapter 4.4). Multivariate extension of such tests and further discussion can be found in the works of Mammen and Nandi (2008) and references therein.

A recent proposal motivated by the method of surrogate data is the TFT-bootstrap of Kirch and Politis (2011). The TFT-bootstrap extends the phase randomization technique since it also resamples the magnitudes of Fourier coefficients and not just their phases. As a result, the scheme is able to correctly capture the distribution of statistics that are based on the periodogram. The authors prove the validity of the scheme in a number of situations, not necessarily linked to testing for nonlinearity.

4. Conclusions

In this chapter, we have presented a selective review on the problem of testing for nonlinearity in time series. Given the broad spectrum of disciplines involved in the exercise some topics have been treated marginally, some have not been mentioned at all. The exposition is informal but aimed at the statistically oriented reader.

The mathematical characterization of linear processes in terms of the Wold theorem and recent results on the (somehow surprising) width of the class of processes implied by such representation indicate that it is virtually impossible to discriminate perfectly between a linear process that satisfies the Wold theorem and any other process. Furthermore, on the side of prediction theory, the optimality of the linear predictor is proved for the class of processes that include the minimum phase ARMA processes. Hence, the linear (Gaussian) paradigm provides us with powerful tools for analyzing time series data but in many instances it fails to capture essential aspects of the process under scrutiny.

Clearly, it is not possible to provide a unified mathematical framework that encompasses all the aspects implied by nonlinearity in the various disciplines. Hence, it is more convenient to test the presence of specific nonlinear features and this is the approach followed in many situations. For instance, tests based on third-order moments and the bispectrum can characterize the nonlinearity linked to the notion of asymmetry and reversibility (see Section 3.1). Initial value sensitivity and nonuniform predictability are assessed through the tests motivated by chaos theory presented in Section 3.5. Also, nonlinearity in the conditional mean or variance can be studied by means of nonparametric regression based tests (Section 3.4). Often, the nonlinearity is assessed in the residuals of a fitted model. This is the approach at the basis of diagnostic tests of Section 3.2. On the contrary, a great deal of literature is devoted to test for the adequacy of a particular nonlinear model, leading to the so called specification tests of Section 3.3. Finally, the surrogate data approach (Section 3.6), introduced in the nonlinear dynamics literature, provides new interesting variations on the topic.

If we compare the problem of testing for nonlinearity in the last 20 years between the fields of Nonlinear Dynamics and Statistics we see that the gap has somehow reduced. In fact, physicists used to deal with very long, often continuous time series, sometimes with low-noise levels and a complex deterministic skeleton. On the contrary, in Statistics, the series are short, very noisy, discrete time and are often the result of aggregation processes. But nowadays, for instance, both time series analysts and physicists can deal with very long (financial) time series in continuous time. Also, some of the notions

motivated by nonlinear dynamics have been adopted also in the time series community. Statisticians, in turn, have provided the necessary mathematical tools for a rigorous approach to data analysis based on such concepts. In this respect, the development of nonparametric Statistics played a central role. The ever increasing availability of computing power can make feasible the application in many Physics- and Engineering-related fields (e.g. EEG or ECG series, meteorology, signal processing, DNA) of the rigorous statistical approach which is very powerful but also time consuming.

It is the opinion of the author that many of the proposals that come from other disciplines, especially nonlinear dynamical system theory, might have a positive impact on the time series community and motivate further research. Clear examples are threshold models (Tong, 2011), tests based on chaos theory, surrogate data methods, and initial value sensitivity in a stochastic environment (Chan and Tong, 2001). Less known examples include resampling methods and MCMC (see, e.g., Mignani and Rosa (2001) for a discussion) and stochastic resonance (Gammaitoni et al., 1998). This latter topic has been touched only marginally by statisticians but might reveal an interesting source of inspiration. In order to be successful, our quest for nonlinearity has to spread across disciplines.

Acknowledgments

I would like to thank Prof. Tata Subba Rao for discussions and comments that led to an improved version of the chapter. This work is dedicated to the loving memory of Paolo Viarengo, brilliant colleague and good friend. The research has been partially supported by MIUR funds.

References

Abarbanel, H., 1996. Analysis of Observed Chaotic Data. Institute for Nonlinear Science. Springer-Verlag, New York.

An, H., Cheng, B., 1991. A kolmogorov-smirnov type statistic with application to test for nonlinearity in time series. Int. Stat. Rev. 59(3), 287–307.

Ashley, R., Patterson, D., Hinich, M., 1986. A diagnostic test for nonlinear serial dependence in time series fitting errors. J. Time Ser. Anal. 7(3), 165–178.

Barnett, A., Wolff, R., 2005. A time-domain test for some types of nonlinearity. IEEE Trans. Signal Process. 53(1), 26–33.

Barnett, W., Gallant, A., Hinich, M., Jungeilges, J., Kaplan, D., Jensen, M., 1997. A single-blind controlled competition among tests for nonlinearity and chaos. J. Econom. 82(1), 157–192.

Berg, A., Paparoditis, E., Politis, D., 2010. A bootstrap test for time series linearity. J. Stat. Plan. Inference 140(12), 3841–3857.

Berg, A., McMurry, T., Politis, D., 2012. Testing time series linearity: traditional and bootstrap methods. In: Rao, C.R., Subba Rao, T. (Eds.), Handbook of Statistics, Volume 30: Time Series. Elsevier, Amsterdam.

Bickel, P., Bühlmann, P., 1996. What is a linear process? Proc. Nat. Acad. Sci. 93, 12128–12131.

Bickel, P., Bühlmann, P., 1997. Closure of linear processes. J. Theoret. Probab. 10(2), 445–479.

Box, G., Jenkins, G., 1970. Time Series Analysis: Forecasting and Control. Holden Day, San Francisco.

Brock, W., Dechert, W., Scheinkman, J., 1986. A test for independence based on the correlation dimension. Econom. Rev. 15(3), 197–235.

Brockett, P., Hinich, M., Patterson, D., 1988. Bispectral-based tests for the detection of gaussianity and linearity in time series. J. Am. Stat. Assoc. 83(403), 657–664.

Brockwell, P., Davis, R., 1991. Time Series: Theory and Methods. Springer-Verlag, New York.

Broer, H., Takens, F., 2011. Dynamical systems and chaos. Vol. 172 of Applied Mathematical Sciences. Springer, New York.

Brooks, C., Hinich, M., 2001. Bicorrelations and cross-bicorrelations as non-linearity tests and tools for exchange rate forecasting. J. Forecast. 20(3), 181–196.

Bühlmann, P., 1997. Sieve bootstrap for time series. Bernoulli 3(2), 123–148.

Chan, K., 1997. On the validity of the method of surrogate data. In: Cutler, C.D., Kaplan, D.T. (Eds.), Nonlinear Dynamics and Time Series. Vol. 11 of Fields Inst. Communications. American Math. Soc., Providence, Rhode Island, pp. 77–97.

Chan, K., Tong, H., 2001. Chaos: A Statistical Perspective. Springer Verlag, New York.

Dahl, C., Gonzlez-Rivera, G., 2003. Testing for neglected nonlinearity in regression models based on the theory of random fields. J. Econom. 114(1), 141–164.

Diks, C., 1999. Nonlinear Time Series Analysis: Methods and Applications. Nonlinear time series and chaos. World Scientific, Singapore.

Dolan, K., Witt, A., Spano, M., Neiman, A., Moss, F., May 1999. Surrogates for finding unstable periodic orbits in noisy data sets. Phys. Rev. E 59(5), 5235–5241.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica 50(4), 987–1007.

Fan, J., Yao, Q., 2003. Nonlinear Time Series. Nonparametric and Parametric Methods. Springer Series in Statistics. Springer-Verlag, New York.

Fan, J., Yao, Q., Tong, H., 1996. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Biometrika 83(1), 189–206.

Fernandes, M., Néri, B., 2010. Nonparametric entropy-based tests of independence between stochastic processes. Econom. Rev. 29(3), 276–306.

Galka, A., 2000. Topics in Nonlinear Time Series Analysis. With Implications for EEG Analysis. Vol. 14 of Advanced Series in Nonlinear Dynamics. World Scientific Publishing Co. Inc., River Edge, NJ.

Gammaitoni, L., Hänggi, P., Jung, P., Marchesoni, F., 1998. Stochastic resonance. Rev. Mod. Phys. 70, 223–287.

Gao, J., 2007. Nonlinear Time Series. Semiparametric and Nonparametric Methods. Vol. 108 of Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, FL.

Giannerini, S., Rosa, R., 2001. New resampling method to assess the accuracy of the maximal Lyapunov exponent estimation. Physica D 155, 101–111.

Giannerini, S., Rosa, R., 2004. Assessing chaos in time series: statistical aspects and perspectives. Stud. Nonlinear Dyn. Econom. 8(2), Article 11.

Giannerini, S., Maasoumi, E., Bee Dagum, E., 2007a. Entropy testing for nonlinearity in time series. In: B. Int. Statist. Inst., 56th session. ISI.

Giannerini, S., Maasoumi, E., Bee Dagum, E., 2011. A powerful entropy test for "linearity" against nonlinearity in time series. working paper.

Giannerini, S., Rosa, R., Gonzalez, D., 2007b. Testing chaotic dynamics in systems with two positive Lyapunov exponents: a bootstrap solution. Int. J. Bifurcat. Chaos 17(1), 169–182.

Granger, C., Andersen, A., 1978. An Introduction to Bilinear Time Series Models. Vandenhoeck & Ruprecht, Göttingen.

Granger, C., Joyeux, R., 1980. An introduction to long-memory time series models and fractional differencing. J. Time Ser. Anal. 1(1), 15–29.

Granger, C., Lin, J., 1994. Using the mutual information coefficient to identify lags in nonlinear models. J. Time Ser. Anal. 15(4), 371–384.

Granger, C., Maasoumi, E., Racine, J., 2004. A dependence metric for possibly nonlinear processes. J. Time Ser. Anal. 25(5), 649–669.

Hannan, E., Deistler, M., 1988. The Statistical Theory of Linear Systems. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

Hinich, M., 1982. Testing for gaussianity and linearity of a stationary time series. J. Time Ser. Anal. 3(3), 169–176.

Hjellvik, V., Tjøstheim, D., 1995. Nonparametric tests of linearity for time series. Biometrika 82(2), 351–368.

Hjellvik, V., Yao, Q., Tjøstheim, D., 1998. Linearity testing using local polynomial approximation. J. Stat. Plan. Inference 68(2), 295–321.

Hong, Y., 1999. Hypothesis testing in time series via the empirical characteristic function: A generalized spectral density approach. J. Am. Stat. Assoc. 94(448), 1201–1220.

Hong, Y., Lee, T., 2003. Diagnostic checking for the adequacy of nonlinear time series models. Econom. Theory 19(6), 1065–1121.

Hong, Y., White, H., 2005. Asymptotic distribution theory for nonparametric entropy measures of serial dependence. Econometrica 73(3), 837–901.
Kantz, H., Schreiber, T., 2004. Nonlinear Time Series Analysis, Second ed. Cambridge University Press, Cambridge.
Kirch, C., Politis, D., 2011. Tft-bootstrap: Resampling time series in the frequency domain to obtain replicates in the time domain. Ann. Stat. 39(3), 1427–1470.
Kugiumtzis, D., 2001. On the reliability of the surrogate data test for nonlinearity in the analysis of noisy time series. Int. J. Bifurcation Chaos 11(7), 1881–1896.
Kugiumtzis, D., Aug 2002. Statically transformed autoregressive process and surrogate data test for nonlinearity. Phys. Rev. E 66(2), 025201.
Kugiumtzis, D., 2008. Evaluation of surrogate and bootstrap tests for nonlinearity in time series. Stud. Nonlinear Dyn. Econom. 12(1), Article 4.
Lee, T.-H., White, H., Granger, C., 1993. Testing for neglected nonlinearity in time series models: a comparison of neural network methods and alternative tests. J. Econom. 56(3), 269–290.
Li, W., 2004. Diagnostic Checks in Time Series. CRC Monographs on Statistics & Applied Probability. Chapman and Hall, Boca Raton, FL.
Lobato, I., 2003. Testing for nonlinear autoregression. J. Bus. Econom. Statist. 21(1), 164–173.
Lorenz, E., 1963. Deterministic nonperiodic flow. J. Atmos. Sci. 20(2), 130–141.
Luukkonen, R., Saikkonen, P., Teräsvirta, T., 1988. Testing linearity in univariate time series models. Scand. J. Stat. 15(3), 161–175.
Maasoumi, E., Racine, J., 2009. A robust entropy-based test of asymmetry for discrete and continuous processes. Econom. Rev. 28(1), 246–261.
Mammen, E., Nandi, S., 2008. Some theoretical properties of phase-randomized multivariate surrogates. Statistics 42(3), 195–205.
Mandelbrot, B.B., 1982. The Fractal Geometry of Nature. W. H. Freeman and Co., San Francisco, California.
Mignani, S., Rosa, R., 2001. Markov Chain Monte Carlo in statistical mechanics: the problem of accuracy. Technometrics 43(3), 347–355.
Milas, C., Rothman, P., van Dijk, D. (Eds.), 2006. Nonlinear time series analysis of business cycles. No. v. 276 in Contributions to economic analysis. Elsevier.
Moran, P., 1953. The statistical analysis of the Canadian Lynx cycle. Aust. J. Zool. 1(3), 291–298.
Park, J., Whang, Y., 2012. Random walk or chaos: A formal test on the Lyapunov exponent. J. Econom. http://dx.doi.org/10.1016/j.jeconom.2012.01.012 (accessed 20.1.12).
Pourahmadi, M., 2001. Foundations of time series analysis and prediction theory. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.
Robinson, P., 1991. Consistent nonparametric entropy-based testing. Rev. Econ. Stud. 58(3), 437–453.
Rosenblatt, M., 2000. Gaussian and Non-Gaussian Linear Time Series and Random Fields. Springer Series in Statistics. Springer-Verlag, New York.
Rusticelli, E., Ashley, R., Bee Dagum, E., Patterson, D., 2009. A new bispectral test for nonlinear serial dependence. Econom. Rev. 28(1), 279–293.
Saikkonen, P., Luukkonen, R., 1988. Lagrange multiplier tests for testing nonlinearities in time series models. Scand. J. Stat. 15(1), 55–68.
Schreiber, T., 1998. Constrained randomization of time series data. Phys. Rev. Lett. 90(10), 2105–2108.
Schreiber, T., Schmitz, A., 1996. Improved surrogate data for nonlinearity tests. Phys. Rev. Lett. 77(4), 635–638.
Schreiber, T., Schmitz, A., 2000. Surrogate time series. Physica D 142(3-4), 346–382.
Shintani, M., Linton, O., 2004. Nonparametric neural network estimation of Lyapunov exponents and a direct test for chaos. J. Econom. 120(1), 1–33.
Small, M., 2005. Applied Nonlinear Time Series Analysis. Applications in Physics, Physiology and Finance. World Scientific, Singapore.
Small, M., Judd, K., 1998. Correlation dimension: A pivotal statistic for non-constrained realizations of composite hypotheses in surrogate data analysis. Physica D 120(3-4), 386–400.
Small, M., Judd, K., Mees, A., 2001. Testing time series for nonlinearity. Stat. comput. 11, 257–268.
Subba Rao, T., 1992. Analysis of nonlinear time series (and chaos) by bispectral methods. In: Casdagli, M., Eubank, S. (Eds.), Nonlinear Modeling and Forecasting. Addison-Wesley, Reading, MA, pp. 199–226.
Subba Rao, T., Gabr, M.M., 1980. A test for linearity of stationary time series. J. Time Ser. Anal. 1(2), 145–158.

Subba Rao, T., Gabr, M.M., 1984. An introduction to bispectral analysis and bilinear time series models. Vol. 24 of Lecture Notes in Statistics. Springer-Verlag, New York.

Subba Rao, T., Wong, W., 1998. Tests for gaussianity and linearity of multivariate stationary time series. J. Stat. Plan. Inference 68(2), 373–386.

Teräsvirta, T., Lin, C., Granger, C., 1993. Power of the neural network linearity test. J. Time Ser. Anal. 14, 209–220.

Terdik, G., Math, J., 1998. A new test of linearity of time series based on the bispectrum. J. Time Ser. Anal. 19(6), 737–753.

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer, J., 1992. Testing for nonlinearity in time series: the method of surrogate data. Physica D 58, 77–94.

Theiler, J., Prichard, D., 1996. Constrained-realization monte-carlo method for hypothesis testing. Physica D 94, 221–235.

Theiler, J., Prichard, D., 1997. Using "surrogate surrogate data" to calibrate the actual rate of false positives in tests for nonlinearity in time series. In: Cutler, C.D., Kaplan, D.T. (Eds.), Nonlinear Dynamics and Time Series. Vol. 11 of Fields Inst. Communications. American Math. Soc., Providence, Rhode Island, pp. 99–113.

Thom, R., 1989. Structural Stability and Morphogenesis. Advanced Book Classics. Addison-Wesley Publishing Company Advanced Book Program, Redwood City, CA.

Tjøstheim, D., 1996. Measures of dependence and tests of independence. Statistics: A J. Theor. Appl. Stat. 28(3), 249–284.

Tong, H., 1990. Nonlinear Time Series. A Dynamical System Approach. Vol. 6 of Oxford Statistical Science Series. The Clarendon Press Oxford University Press, New York, with an appendix by K. S. Chan, Oxford Science Publications.

Tong, H., 2011. Threshold models in time series analysis: 30 years on. Stat. Interfac. 4(2), 107–118.

Tsay, R., 2005. Analysis of Financial Time Series. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ.

Whang, Y.-J., Linton, O., 1999. The asymptotic distribution of nonparametric estimates of the Lyapunov exponent for stochastic time series. J. Econom. 91(1), 1–42.

Yao, Q., Tong, H., 1994a. On prediction and chaos in stochastic systems. Philos. Trans. R. Soc. Lond. A 348, 357–369.

Yao, Q., Tong, H., 1994b. Quantifying the influence of initial values on non-linear prediction. J. R. Stat. Soc. B 56, 701–725.

This page intentionally left blank

Part II: Nonlinear Time Series

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00004-1

Modelling Nonlinear and Nonstationary Time Series

Dag Tjøstheim

Department of Mathematics, University of Bergen, Johs. Brunsgt. 12, 5008 Bergen, Norway

Abstract

An overview is given of the modelling of nonlinear and nonstationary time series. The emphasis is on the theory for time series that are both nonlinear and nonstationary. But to put that topic into perspective, brief outlines of the theory for nonlinear and stationary and for linear and nonstationary models are also given. Topics such as nonlinear integrated processes and nonlinear cointegrating regression are included, and both parametric and nonparametric estimation are considered.

Keywords: nonlinear time series, nonstationary time series, nullrecurrent Markov chain, nonlinear integrated process, nonlinear cointegrating regression, nonparametric estimation.

1. Introduction

The concepts of linearity and stationarity give rise to a pair of two way classifications of time series models: linear or nonlinear and stationary or nonstationary. Most model used in practice have been linear and stationary. However, in the last three or four decades, there has been a lot of interest in linear models that have a particular, somewhat narrow, form of nonstationarity, the unit root processes. This has been followed up for multivariate linear processes resulting in the important class of cointegrating models (Engle and Granger, 1987; Johansen, 1995; Juselius, 2006). Parallel to this, there has been a strong development in nonlinear stationary models (Fan and Yao, 2003; Tong, 1990), where, e.g., threshold models and smooth transition autoregressive (STAR) models are now being extensively used. The final category, which is by far the largest and is also the least studied, consists of nonlinear nonstationary processes. This

class of processes will be the main object of interest in this survey, but to put matters into perspective, we will start by looking at nonlinear stationary processes in Section 2 and at linear nonstationary processes in Section 3.

Our approach will be virtually entirely in time domain. There are nonlinear models in frequency domain, but on the whole, time domain models have been much more fruitful at least in applications to finance and economics. Actually, in these areas, it is not uncommon to have time series that could be both nonstationary and nonlinear.

The extension of the estimation and specification theory to nonlinear and nonstationary models involves hard mathematical challenges. In the linear stationary case, mixing and martingale theory can be used to obtain distributional limit results. In the nonlinear stationary situation, these tools can still be used, but the issue of actually determining whether a given model is stationary becomes important and difficult. Markov chain theory is an additional device for this purpose as well as for the asymptotic theory. Still, in this situation, asymptotic normality of estimators is the rule. This changes in the linear nonstationary case, where non-Gaussian distributions and functions of Brownian motion play a far greater role. For models that are both nonlinear and nonstationary, again new concepts are needed, such as local time and null recurrence of Markov chains. These concepts can be expected to play a role in attempts to construct a nonlinear cointegration theory. Existing results in these areas are few, and difficult problems remain to be solved.

Both parametric and nonparametric estimation will be looked at in this survey. Parametric models are sometimes easier to treat theoretically, and the convergence rate is typically faster, although not necessarily as fast as in the linear unit root case. A disadvantage is of course that a given parametric model may not be appropriate for the data at hand. In this sense, a nonparametric approach is more flexible. The price paid for larger flexibility is a slower convergence rate and, in some cases, a more complicated theory. In the framework of nonlinearity and nonstationarity, parametric models are treated in Sections 4.2–4.4, whereas the nonparametric approach will be reviewed in Section 4.5. The type of nonstationarity that we will concentrate on will be the nonlinear random walk type, where in the parametric case, the parameters are fixed in time. Another possibility is to consider models (linear and nonlinear) with time-varying parameters often in a state-space setting. This strand of the literature will be briefly mentioned in Section 5. Much of our material is taken from Teräsvirta et al. (2010) to which we refer for more details and related topics.

2. Nonlinear stationary models

For a process $\{y_t\}$ with $t \geq 0$ or $-\infty < t < \infty$, strict stationarity is defined by requiring that the joint distribution of $(y_{t_1}, y_{t_2}, \ldots, y_{t_k})$ is the same as that of $(y_{t_1+t}, y_{t_2+t}, \ldots, y_{t_k+t})$ for every t and every set of time points $(t_1, t_2, \ldots, t_k)$ in the domain of definition. It is much more difficult to find a precise definition of nonlinearity. This is illustrated by the following example where one starts with an obvious example of a linear process, namely a moving-average, MA(q), process:

$$y_t = \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

where $\{\varepsilon_t\} \sim$ i.i.d.$(0, \sigma^2)$ (identically independently distributed with mean 0 and variance σ^2 and $\{\theta_j, j = 1, \ldots, q\}$ are any set of weights. It has been found that the closure (that is, the set of MA(q) processes as $q \to \infty$ in the so-called Mallows topology) has rather complicated properties and can include many processes that one would think of as highly nonlinear; see Bickel and Bühlmann (2003). The result is interesting and highly technical, but will not be discussed more here. Instead, we take a more pragmatic point of view and mainly think of a nonlinear model as that defined by a nonlinear possibly nonhomogenous autoregressive model given by

$$y_t = g(\mathbf{y}_{t-1}, \mathbf{x}_t) + h(\mathbf{y}_{t-1}, \mathbf{x}_t)\varepsilon_t, \quad t \geq t_0 \tag{1}$$

where g and h are functions, where $\mathbf{y}_{t-1} = (y_{t-1}, \ldots, y_{t-p})'$ consists of lags of y_t, and $\mathbf{x}_t = (x_{1t}, \ldots, x_{kt})'$ is a vector of exogenous variables.

2.1. Stationarity of nonlinear models

Whereas the probability structure of linear Gaussian models is determined by second-order moments, for nonlinear models, moments will not in general suffice. One needs to look at the distribution and at conditional quantities such as the conditional mean and the conditional variance. This means that strict sense stationarity rather than wide sense stationarity is required. The definition given of stationarity in the beginning of this section is in a sense too strict and in another sense too wide. It is too strict because we need to allow for processes that are asymptotically stationary but can be started with an arbitrary initial condition. It is too wide because stationarity in itself is usually not enough to establish limit theorems. What is needed is some form of ergodicity, and where the Markov property could be quite essential.

We illustrate this by the scalar recursively defined Markov model $\{y_t, t \geq 1\}$ modelled by,

$$y_t = g(y_{t-1}, \theta) + \varepsilon_t, \quad y_0 = y(0) \tag{2}$$

where g can be taken as a known function, θ as an unknown parameter, and $\{\varepsilon_t\} \sim$ i.i.d.$(0, \sigma^2)$ and ε_t independent of $\{y_s, s < t\}$. The stationarity criteria to be treated below furnish existence result in the sense that they guarantee the existence of an initial distribution for y_0 such that the system is stationary when started with this distribution. If it is started with another distribution, the effect of this initial distribution will die out eventually, so that the process is "asymptotically stationary" or "stable." In addition, these criteria also imply ergodicity, which is one tool that can be used for establishing asymptotics.

A simple special case is the linear Gaussian autoregressive process

$$y_t = \phi y_{t-1} + \varepsilon_t, \quad y_0 = y(0)$$

where, if $|\phi| < 1$ and $y(0)$ is normally distributed with zero mean and variance $\sigma^2(1 - \phi^2)^{-1}$, the process $\{y_t\}$ is strictly stationary (and all moments exist). If it is started with another distribution, e.g., $y(0)$ could be a fixed number, then

$$y_t = \phi^t y(0) + \sum_{i=0}^{t-1} \phi^i \varepsilon_{t-i}.$$

The effect from the initial condition will die out since $|\phi| < 1$, and $\{y_t\}$ will approach the stationary process defined by

$$y_t = \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i}.$$

with $\{\varepsilon_t, -\infty < t < \infty\} \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$. For a nonlinear system, a corresponding condition is far more difficult to find, and appropriate assumptions is often best discussed for specific models. Suffice here to say is that if ε_t has a continuous distribution, if the function $g(y, \theta)$ in (2) is bounded for y in a compact set, and if there exists a non-negative test function V satisfying a Foster-Lyapunov drift criterion, i.e.,

$$\mathsf{E}\{V(y_t)|y_{t-1} = y\} = \mathsf{E}\{V(g(y, \theta) + \varepsilon_t)\} < V(y)$$

for large values of $|y|$, then an initial distribution can be found such that $\{y_t\}$ is (strictly) stationary. Taking $V(y) = |y|^2$, a sufficient condition is that $|g(y, \theta)| < |y|$ for $|y|$ large, and if $\mathsf{E}\varepsilon_t^2 < \infty$, the existence of second-order moments is also assured. This means that if the nonlinear dynamics is bounded far out by a stationary linear autoregressive process, then a stationary solution exists. Actually, such an approach often yields a great deal more in that so-called geometric ergodicity is obtained, which is very useful in proving central limit theorems in estimation. We refer to Meyn and Tweedie (1993) for much more details and similar results.

In general, the problem of finding conditions for stationarity is difficult, especially for the nonlinear higher-order AR case. Even for a second-order threshold AR process a complete characterization does not seem to be known. The Foster-Lyapunov criterion can also be used for ARCH/GARCH process. Recent publications in this area include Carrasco and Chen (2002), Francq and Zakoïan (2006), Liebscher (2005), Ling and McAleer (2002), Meitz and Saikkonen (2008), and Meitz and Saikkonen (2011). For all types of Markov models, much effort has been spent on proving geometric ergodicity, since this implies mixing, which in turn can be used for proving central limit theorems.

An alternative to the Markov chain approach is to view (2) as a stochastic recursive equation and use a contraction principle to find conditions for existence of stationary solutions. This way of attacking the problem has proved especially useful for nonlinear conditional heteroskedastic models. We refer to Straumann (2005) and Aue et al. (2006).

For some nonstationary time series, it may be possible to transform them to stationarity to take advantage of the theory for stationary models. In the linear univariate case such a transformation would typically consist in differencing the series. The class of series for which differencing achieves stationarity is closely related to the class of I(1) and I(2) (differencing twice) processes, or the unit root processes (cf. Section 3.1 and Hamilton, 1994, Chapters 17 and 19). In the multivariate linear case, cointegration theory can be used to find linear transformations to a system of jointly stationary processes (cf. Section 3.2 and Hamilton, 1994, Chapter 19).

In the nonlinear nonstationary case, again all this becomes more difficult. As in the linear case, restrictions on the type of nonstationarity allowed have to be introduced. There are presently two possibilities, first, to use decomposition in terms of

Brownian motion processes with local time, and, second, to apply recurrence theory of Markov chains. Both approaches are highlighted in Section 4. This is a fairly new area of research and one with a great deal of potential, at least for theoretical development.

In Karlsen and Tjøstheim (2001) and Karlsen et al. (2007), nonlinear unit root processes and nonlinear cointegration type models are analyzed in terms of null recurrent Markov chains, see Sections 4.1 and 4.5. It then becomes of interest to find conditions on $g(y, \boldsymbol{\theta})$ which imply null recurrence. This work has just started, see Myklebust et al. (2011) which is mostly about the linear vector case, though. As an example of the difficulties involved, it is known that the threshold-like process with random walk behavior far out,

$$y_t = g(y_{t-1}, \boldsymbol{\theta}) I(|y_{t-1}| \leq c) + y_{t-1} I(|y_{t-1}| > c) + \varepsilon_t$$

where $I(\cdot)$ is the indicator function and is null recurrent if $\sup_{|y| \leq c} g(y, \boldsymbol{\theta}) < \infty$, but to our knowledge, it is not known whether the exponential AR process

$$y_t = (1 + \psi e^{-\gamma y_{t-1}^2}) y_{t-1} + \varepsilon_t, \quad \gamma > 0$$

is null recurrent. Intermediate cases between geometric ergodicity and null recurrence have received attention as well (cf. Yao and Attali, 2000).

2.2. *Some specific nonlinear models*

In Section 2.1, we took linear autoregressive models as a prototype of linear models to start from. One may wonder why not starting with moving-average models. In one sense, the moving-average processes have the simplest mathematical structure, but little progress has been made so far with nonlinear moving-average models. Moreover, the apparently simpler mathematical structure of moving-average models could be deceptive. This is very effectively illustrated by Breidt et al. (2006) for the linear MA(1) model. Explicit nonlinear extensions of the moving-average model do exist; see, for example, de Gooijer (1998), Ling and Tong (2005), and Li et al. (2011) for a threshold moving-average model, but the theory has not progressed very far. It seems that when it comes to formulating nonlinear extensions that are useful in practice, starting with a linear autoregressive model is far preferable, primarily because their recursive structure makes them easier to estimate and forecast. In this section, we shall mainly discuss some specific extensions of linear autoregressive processes.

Virtually all of the important aspects of nonlinearity emerge in the first-order case, and to begin with, we will therefore concentrate on generalizations of the linear AR(1) model

$$y_t = \phi y_{t-1} + \varepsilon_t, \tag{3}$$

where ϕ is the autoregressive coefficient and $\{\varepsilon_t\} \sim \text{i.i.d.}(0, \sigma^2)$.

Aspects of the estimation theory for a parametric nonlinear AR process

$$y_t = g(y_{t-1}, \boldsymbol{\theta}) + \varepsilon_t$$

are already covered in Tjøstheim (1986), but there are many more recent contributions, and the reader is referred to Fan and Yao (2003, Chapter 4) and Ling and McAleer (2010).

Another generalization of (3) is to consider a model

$$y_t = \phi y_{t-1} + h(y_{t-1})\varepsilon_t, \tag{4}$$

where h may be unknown or known up to a parameter. Such a model can be modified and extended in many directions to obtain models of autoregressive conditional heteroskedasticity that have been considered in Teräsvirta et al. (2010, Chapter 8). By again replacing ϕy_{t-1} by $g(y_{t-1})$ in (4), one obtains a model that is nonlinear both in the conditional mean and the conditional variance but fairly difficult to specify and estimate.

A rather different generalization of (3) is obtained by replacing ϕ by a random process $\{\theta_t\}$, where $\{\theta_t\}$ may itself be an autoregressive process. This can be taken as an example of a state-space process, and it will be mentioned briefly in Section 5.

The classic nonlinear parametric time series models include the threshold autoregressive (TAR) model (Tong and Lim, 1980; Tong, 1990), the exponential autoregressive model (Ozaki, 1982, 1985) and the bilinear model (Subba Rao, 1981; Subba Rao and Gabr, 1984). It is probably fair to say that of these three models, the threshold model has been found to be most useful in practice. In its simplest form, the linear model (3) is replaced by a nonlinear mechanism such that for every t, y_t is generated by one of two linear models. The value of y_{t-1}, the threshold variable, determines which model. More formally,

$$y_t = \phi_1 y_{t-1} I(y_{t-1} \le c) + \phi_2 y_{t-1} I(y_{t-1} > c) + \varepsilon_t$$

where c is the threshold parameter. In general, the threshold variable could be chosen to be another time lag, y_{t-d}, say. There are numerous generalizations, not necessarily restricted to the stationary case. For recent applications of the threshold model to nonlinear error correction models, see Hansen and Seo (2002), Hansen (2003), Bec and Rahbek (2004), and Section 4.3. The asymptotic theory in the stationary case can be found in Chan (1993) and in Li and Ling (2011). The maximum likelihood estimator of the threshold parameter has a nonstandard asymptotic distribution and has a faster convergence rate.

The TAR model has been criticized for its lack of smoothness in its transition mechanism. The exponential autoregressive model given by

$$y_t = (\phi + \psi e^{-\gamma y_{t-1}^2}) y_{t-1} + \varepsilon_t,\ \gamma > 0$$

was introduced partly as a response to this criticism. For large $|y_{t-1}|$, the process essentially moves according to an AR process with parameter ϕ, whereas for small $|y_{t-1}|$, the time-varying AR coefficient is roughly $\phi + \psi$. This is an example of a smooth transition model on which there is a rich literature, see e.g., Teräsvirta et al. (2010).

Many of the models treated above have trivial extensions to the higher-order case and to vector models. As an example, we mention the vector threshold model and the smooth transition vector autoregressive model.

The vector threshold autoregressive model can be defined as Tsay (1998),

$$\mathbf{y}_t = \sum_{i=1}^{r} \left\{ \sum_{j=1}^{p} (\boldsymbol{\mu}_i + \Phi_{ij}\mathbf{y}_{t-j} + \Gamma_{ij}\mathbf{x}_{t-j}) + \boldsymbol{\varepsilon}_{it} \right\} I(c_{i-1} < s_t \leq c_i) \tag{5}$$

where $\mathbf{y}_t$ and $\boldsymbol{\varepsilon}_{it}$, $i = 1, \ldots, r$, are stochastic $m \times 1$ vectors and $\boldsymbol{\mu}_i$ is an $m \times 1$ vector of intercepts, $i = 1, \ldots, r$. Furthermore, Φ_{ij} are $m \times m$ coefficient matrices and Γ_{ij} are $m \times k$ coefficient matrices, both for $i = 1, \ldots, r$, and $j = 1, \ldots, p$. The errors $\boldsymbol{\varepsilon}_{it}$ are serially uncorrelated with mean $\mathbf{0}$ and positive definite covariance matrices $\Sigma_i, i = 1, \ldots, r$. A single stationary and continuous switch-variable s_t determines the regime of the whole system, where $\mathbf{x}_t$ is a time series of explanatory variables. A modelling strategy for this family of models with applications is developed and discussed in Tsay (1998).

The exponential AR model was mentioned as an example of a smooth transition autoregressive model. The idea of a smooth transition may be generalized to the vector case. A vector STAR model of order p may be defined as follows:

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{\mu}_0 + \sum_{j=1}^{p} \Phi_j \mathbf{y}_{t-j} + \mathbf{G}(\boldsymbol{\gamma}, \mathbf{c}; \mathbf{s}_t) \left(\boldsymbol{\mu}_1 + \sum_{j=1}^{p} \Psi_j \mathbf{y}_{t-j} \right) + \boldsymbol{\varepsilon}_t \\ &= \boldsymbol{\mu}_0 + \mathbf{G}(\boldsymbol{\gamma}, \mathbf{c}; \mathbf{s}_t)\boldsymbol{\mu}_1 + \sum_{j=1}^{p} \{\Phi_j + \mathbf{G}(\boldsymbol{\gamma}, \mathbf{c}; \mathbf{s}_t)\Psi_j\}\mathbf{y}_{t-j} + \boldsymbol{\varepsilon}_t, \end{aligned}$$

where $\mathbf{y}_t$ is an $m \times 1$ vector, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are $m \times 1$ intercept vectors, Φ_j and Ψ_j, $j = 1, \ldots, p$, are $m \times m$ parameter matrices, and

$$\mathbf{G}(\boldsymbol{\gamma}, \mathbf{c}; \mathbf{s}_t) = \text{diag}\{G_1(\gamma_1, \mathbf{c}_1, s_{1t}), \ldots, G_m(\gamma_m, \mathbf{c}_m, s_{mt})\}$$

is the $m \times m$ diagonal matrix of transition functions. Furthermore, $\boldsymbol{\varepsilon}_t \sim$ i.i.d. $\mathcal{N}(\mathbf{0}, \Sigma)$, where $\boldsymbol{\Sigma}$ is positive definite.The logistic vector STAR (LVSTAR) results if $G_j(\gamma_j, \mathbf{c}_j, s_{jt})$, $j = 1, \ldots, m$, are standard logistic functions; i.e,

$$G(\gamma, \mathbf{c}, s_t) = \left(1 + \exp\left\{ -\gamma \prod_{k=1}^{K} (s_t - c_k) \right\} \right)^{-1}, \qquad \gamma > 0$$

where $\gamma > 0$ is an identifying restriction.

Much more about STAR models can be found in Teräsvirta et al. (2010).

In spite of the large flexibility of STAR models, we end this section by briefly mentioning two other model classes, one parametric, the artificial neural network (ANN) models, and the other nonparametric, the additive models.

ANN models are worth mentioning because of the attention they have received both in statistical and neural network literature and among practicians. In the simplest single-equation case, which is the so-called "single hidden-layer" model, it has the form

$$y_t = \boldsymbol{\beta}_0' \mathbf{z}_t + \sum_{j=1}^{q} \beta_j G(\boldsymbol{\gamma}_j' \mathbf{z}_t) + \varepsilon_t \tag{6}$$

where y_t is the output series, $\mathbf{z}_t = (1, y_{t-1}, \ldots, y_{t-p}, x_{1t}, \ldots, x_{kt})'$ is the vector of inputs, including the intercept and lagged values of the output, $\boldsymbol{\beta}_0'\mathbf{z}_t$ is a linear unit with $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \ldots, \beta_{0,p+k})'$. Furthermore, β_j, $j = 1, \ldots, q$, are parameters, called "connection strengths" in the neural network literature. The function $G(\cdot)$ is a bounded, asymptotically constant function called the "squashing function" and $\boldsymbol{\gamma}_j$, $j = 1, \ldots, q$, are parameter vectors. Typical squashing functions are monotonically increasing ones such as the logistic function and the hyperbolic tangent function. The errors ε_t are often assumed i.i.d.$(0,\sigma^2)$. The term "hidden layer" refers to the structure of (6). While the output y_t and the input vector $\mathbf{z}_t$ are observable, the linear combination $\sum_{j=1}^{q} \beta_j G(\boldsymbol{\gamma}_j'\mathbf{z}_t)$ is not. It thus forms a hidden layer between the "output layer" y_t and "input layer" $\mathbf{z}_t$.

The use of ANN in building time series models is discussed in Chapter 16 of Teräsvirta et al. (2010).

ANN models are almost nonparametric in structure, and at this point, it is also useful to mention a general type of higher-order model, the class of additive models that are especially useful for nonparametric analysis. A general regression model commonly assumed in economics is

$$y_t = g(\mathbf{z}_t) + h(\mathbf{z}_t)\varepsilon_t$$

where $\mathbf{z}_t$ is a vector of explanatory variables, including lags of y_t and where ε_t is assumed to be independent of $\mathbf{z}_t$. With known forms for the functions g and h except for certain parametric values, maximum likelihood methods can efficiently estimate these parameters if the assumptions of the model are appropriate. A more realistic situation is when g and h may be unknown functions. In a few simple cases, with a low-dimensional $\mathbf{z}_t$, they can be estimated nonparametrically, but many interesting models cannot be handled, primarily because of the curse of dimensionality when the dimension of $\mathbf{z}_t$ exceeds 3-4.

A useful approximation, in the high-dimensional case, is to consider a simple additive model of the form

$$g(\mathbf{z}_t) = g_0 + \sum_{j=1}^{m} g_j(z_{jt})$$

where there are m explanatory variables. As Sperlich et al. (2002) point out, such models have a long and distinguished history in both economics and statistics.

However, one is often interested in interactions, for instance in economics, so Sperlich et al. (2002) consider a wider class of models, where

$$g(\mathbf{z}_t) = g_0 + \sum_{j=1}^{m} g_j(z_{jt}) + \sum_{1<i<j\leq m} g_{ij}(z_{it}, z_{jt})$$

so that the completely general function g is approximated by functions of single z_{jt}'s and pairs of z_{jt}'s. In most "well behaved" circumstances, the approximation can be thought of as likely to be an acceptable one.

3. Linear nonstationarity

To put the models of the next section into perspective, we include a brief review of linear processes which are nonstationary. The nonstationarity is of random walk type. These models have been much used by econometricians.

3.1. Linear unit root models

A natural starting point for discussing nonstationary models is the unit root model which takes the form

$$(1 - \mathsf{L})y_t = \Delta y_t = x_t \tag{7}$$

where L is the backshift operator and $\{x_t\}$ is a stationary series with constant variance σ_x^2. If the series begins at $t = 0$, the variance of y_t is approximately $t\sigma_x^2$ and, as this variance changes with time, $\{y_t\}$ is nonstationary. Of course, such a process merely represents a simple case of nonstationarity and is hardly a typical example of this category, even though it is sometimes used as such. If $\{x_t\}$ has a nonzero mean, $\{y_t\}$ will also have a linear trend in mean. An example of a unit root process is the simple random walk

$$\Delta y_t = \varepsilon_t \tag{8}$$

where $\{\varepsilon_t\} \sim$ i.i.d.(μ, σ^2). Then, y_t has mean μt and variance $\sigma^2 t$.

In the common notation used in this subject, a series is denoted as I(1) if it has to be differenced once to achieve stationarity and I(2) if it has to be differenced twice for that. One might think that a stationary process may be reasonably defined as I(0). Unfortunately, this latter definition is not quite strict enough as the fractional ARMA processes are denoted by I(d) and are stationary for $0 < d < 1/2$. A better definition in the linear framework is that an I(0) series is ARMA with spectrum bounded above and away from zero at all frequencies. We will return to the problem of defining I(0) in the nonlinear case in Section 4.1.

A rather complete theory of statistical inference is now available for linear unit root models; see e.g., Phillips (1987), Phillips and Solo (1992), and Tanaka (1996). It is worthwhile to go through the crucial steps of such an analysis to highlight the contrasts with the nonlinear case to be treated in Section 4.4. The essential features of the analysis already emerge in the simple random walk case. Consider the random walk model (8). Its stationary counterpart is the AR(1) process

$$y_t - \phi y_{t-1} = \varepsilon_t$$

with $|\phi| < 1$. In the latter case, if $\mathsf{E}y_t = 0$,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_t \xrightarrow{d} \mathcal{N}(0, \sigma_y^2)$$

and for the least squares estimator (which equals the maximum likelihood estimator asymptotically)

$$\widehat{\phi} = \frac{\sum y_t y_{t-1}}{\sum y_t^2}$$

of ϕ we have

$$\sqrt{T}(\widehat{\phi} - \phi) = \frac{T^{-1/2} \sum \varepsilon_t y_{t-1}}{T^{-1} \sum y_t^2}.$$

Furthermore, a central limit theorem is obtained from the fact that

$$T^{-1} \sum y_t^2 \xrightarrow{p} \sigma_y^2 \quad \text{and} \quad T^{-1/2} \sum \varepsilon_t y_{t-1} \xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2 \sigma_y^2\right).$$

where $\xrightarrow{p}$ and $\xrightarrow{d}$ denote convergence in probability and distribution, respectively. When $\phi = 1$ as in (8), everything changes dramatically. The limit distribution of $\widehat{\phi}$ in this case can be derived using the so-called functional limit theorem (also called the invariance principle) as is described in e.g., Tanaka (1996, Chapter 3). Because of the comparison to the nonlinear case, we now give a brief heuristic derivation of asymptotic results for $\{y_t\}$ in terms of the Wiener process. A standard Wiener process (or Brownian motion) is a Gaussian stochastic process $\{w_t\}$ in continuous time with independent increments such that $w_0 = 0$, $\mathsf{E}w_t = 0$, and $\operatorname{corr}(w_t, w_s) = \min(t, s)$.

Assume that $\{y_t\}$ is a random walk process as in (8). If $\sigma_\varepsilon = 1$, we have from the classic central limit theorem that

$$\frac{y_T}{\sqrt{T}} = \frac{1}{\sqrt{T}} \sum_{s=1}^{T} \varepsilon_s \xrightarrow{d} w_1 \sim \mathcal{N}(0, 1). \tag{9}$$

Similarly, it can be shown that for $0 \le r \le 1$,

$$\frac{y_{[rT]}}{\sqrt{T}} = \frac{1}{\sqrt{T}} \sum_{s=1}^{[rT]} \varepsilon_s \xrightarrow{d} w_r \tag{10}$$

where $[rT]$ is the integer part of rT defined as the largest integer less than or equal to rT. It follows heuristically from (10); see e.g., Tanaka (1996, Chapter 3) for easy but rigorous proofs, that

$$\frac{1}{T^2} \sum_{t=1}^{T} y_t^2 = \sum_{t=1}^{T} \left(\frac{y_t}{\sqrt{T}}\right)^2 \cdot \frac{1}{T} \xrightarrow{d} \int_0^1 w_r^2 dr \tag{11}$$

where $1/T$ plays the role of the differential dr when T gets large. Similarly,

$$\frac{1}{T} \sum \varepsilon_t y_{t-1} = \sum_{t=1}^{T} \frac{y_{t-1}}{\sqrt{T}} \cdot \frac{y_t - y_{t-1}}{\sqrt{T}} \xrightarrow{d} \int_0^1 w_r dw_r \tag{12}$$

where the integral is interpreted as an Itô-integral. Note that in contrast to the stationary case, in (11) (and in (12)), one does not have convergence to a number but to a stochastic variable.

The function $h(y) = \int_0^1 y_t^2 dt$ is continuous on the space C of continuous functions on [0,1], and the continuous mapping theorem is used to prove (11) and with some modifications (12). Moreover, it can be used to prove (Tanaka, 1996, p. 75) that for model (8) with $\widehat{\phi} = \sum y_t y_{t-1} / \sum y_t^2$,

$$T(\widehat{\phi} - 1) = \frac{\frac{1}{T}\sum \varepsilon_t y_{t-1}}{\frac{1}{T^2}\sum y_t^2} \xrightarrow{d} \frac{\int_0^1 w_r dw_r}{\int_0^1 w_r^2 dr}$$

which shows that $\widehat{\phi}$ is super consistent as an estimator of $\phi = 1$. This result is used as a point of departure for unit root tests of Dickey-Fuller type.

It should be noted that the above results can be derived (cf. Tanaka 1996, Chapter 1) without functional limit arguments and the Wiener process. However, the functional limit theory yields a very powerful instrument for generalizing the results, so that they cover (7) for quite general processes $\{x_t\}$. Thus, for (8), the results can be generalized (Phillips, 1987) to a situation where $\{\varepsilon_t\}$ is replaced by a process $\{x_t\}$ which is a martingale process. A mixing process and even heteroskedasticity may be allowed. With these generalizations, $y_{[rT]}/\sqrt{T} \xrightarrow{d} w_r$ still holds. The expression $\sum y_t^2$ that appears in the linear estimation problem already involves a nonlinear transformation of an I(1) process. However, for nonlinear estimation theory of I(1) processes and other nonstationary processes, more complicated nonlinear transformations are required. We shall return to this point in Section 4.1 but shall first present a few aspects of the theory for linear vector models.

3.2. *Vector autoregressive processes and linear cointegration*

The full significance of the I(1) concept can first be realized in a joint description of a vector time series. Such systems of time series are often modelled by a vector AR (VAR) process

$$\mathbf{y}_t = \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t = \sum_{i=1}^{p} \mathbf{\Phi}_i \mathsf{L}^i \mathbf{y}_t + \boldsymbol{\varepsilon}_t \tag{13}$$

where $\mathbf{y}_t$ is m-dimensional and the matrices $\mathbf{\Phi}_i$ are $m \times m$. The vector time series $\{\mathbf{y}_t\}$ is stationary if the roots of the characteristic polynomial $\mathbf{\Phi}(z) = I_m - \sum_{i=1}^{p} \mathbf{\Phi}_i z^i$ are outside the unit circle, that is, if $|\mathbf{\Phi}(z)| \neq 0$ for $|z| \leq 1$. Here, $\mathbf{I}_m$ is the m-dimensional identity matrix. If there are k unit roots, say, and $m - k$ roots outside the unit circle, $\{\mathbf{y}_t\}$ is nonstationary and the components are I(1) or I(0). In the trivial and completely uninteresting case of independence between the component processes there are exactly k I(1) processes and $m - k$ I(0) processes. In the case of dependence between the component processes, the k unit roots correspond to k common stochastic trends, and the $m - k$ roots outside the unit circle lead to the existence of $m - k$ linear combinations (eigenvectors corresponding to these roots) of the components which are stationary I(0)

even though the component processes are nonstationary I(1). This property is called cointegration. The processes making up each of the $m-k$ linear combinations move together in the long run. The cointegration concept was introduced in Granger (1981) and further developed in Engle and Granger (1987) and has spawned numerous papers.

There are two main representations of a cointegrated system, the error correction representation and the triangular representation. Both of these have served as a basis for nonlinear extensions. The error correction representation is obtained by subtracting $\mathbf{y}_{t-1}$ from both sides of (13) and rearranging this equation as

$$\Delta \mathbf{y}_t = \mathbf{C}\mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \boldsymbol{\Psi}_i \Delta \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t \tag{14}$$

where $\mathbf{C} = -\mathbf{I}_m + \sum_{i=1}^{p} \boldsymbol{\Phi}_i = -\boldsymbol{\Phi}(1)$ and $\boldsymbol{\Psi}_i = -\sum_{j=i+1}^{p} \boldsymbol{\Phi}_j, i = 1, \ldots, p-1$. When there are k unit roots of the characteristic polynomial, the matrix $\mathbf{C} = -\boldsymbol{\Phi}(1)$ has rank $n = m - k$. The row space of $\mathbf{C}$ is then spanned by a basis of n linearly independent vectors, and we denote by $\boldsymbol{\alpha}$ the $m \times n$ matrix whose columns form such a basis. Every row of $\mathbf{C}$ can now be written as a linear combination of the rows of $\boldsymbol{\alpha}'$. Thus, we can write $\mathbf{C} = \boldsymbol{\delta}\boldsymbol{\alpha}'$, where $\boldsymbol{\delta}$ is an $m \times n$ matrix with full column rank, and Eq. (14) can then be written

$$\Delta \mathbf{y}_t = \boldsymbol{\delta}\boldsymbol{\alpha}'\mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \boldsymbol{\Psi}_i \Delta \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t$$

or

$$\Delta \mathbf{y}_t = \boldsymbol{\delta}\mathbf{x}_{t-1} + \sum_{i=1}^{p-1} \boldsymbol{\Psi}_i \Delta \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t \tag{15}$$

where $\mathbf{x}_{t-1} = \boldsymbol{\alpha}'\mathbf{y}_{t-1}$. One can solve for $\mathbf{x}_{t-1}$ obtaining

$$\mathbf{x}_{t-1} = (\boldsymbol{\delta}'\boldsymbol{\delta})^{-1}\boldsymbol{\delta}' \left[\Delta \mathbf{y}_t - \sum_{i=1}^{p-1} \boldsymbol{\Psi}_i \Delta \mathbf{y}_{t-i} - \boldsymbol{\varepsilon}_t \right] \tag{16}$$

so that $\mathbf{x}_t$ is I(0). Thus, the linear combinations $\mathbf{x}_t = \boldsymbol{\alpha}'\mathbf{y}_t$ of nonstationary components are stationary, and the columns of $\boldsymbol{\alpha}$ are the cointegrating vectors. The term "error correction" first appeared in Phillips (1957) and another pioneer was Sargan (1964). The relationship $\boldsymbol{\alpha}'\mathbf{y}_t = \mathbf{0}$ can be interpreted as the "equilibrium" of the dynamical system and $\mathbf{x}_t$ as the vector of "equilibrium errors" and Eq. (16) then describes the self correcting mechanism of the system. The error correction representation has been further developed in several papers, see e.g., Johansen (1988, 1991, 1995). The basis of these developments in the statistical literature is reduced rank regression.

The other representation of a cointegrated VAR system is based on a matrix polynomial decomposition $\boldsymbol{\Phi}(z) = \mathbf{U}(z)\mathbf{M}(z)\mathbf{V}(z)$. Here, $\mathbf{U}(z)$ and $\mathbf{V}(z)$ are $m \times m$ matrix

polynomials with all their roots outside the unit circle, and $\mathbf{M}(z)$ is a $m \times m$ diagonal matrix polynomial with roots on or outside the unit circle. In the case of a cointegrated VAR, $\mathbf{M}(z)$ can be written as follows:

$$\mathbf{M}(z) = \begin{bmatrix} \boldsymbol{\Delta}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix}$$

where $\Delta_k = (1 - z)\mathbf{I}_k$. Hence all of the nonstationarity of the VAR is in the upper block of $\mathbf{M}(z)$. Using this decomposition and a rearrangement of the components (cf. Watson, 1994, pp. 2872–4) of $\mathbf{y}_t$ implies that $\mathbf{y}_t$ can be written as $\mathbf{y}_t = (\mathbf{y}'_{1t}, \mathbf{y}'_{2t})'$, where $\mathbf{y}_{1t}$ is k-dimensional and $\mathbf{y}_{2t}$ is $n = (m - k)$-dimensional with the triangular representation

$$\begin{aligned} \Delta\mathbf{y}_{1t} &= \mathbf{u}_{1t} \\ \mathbf{y}_{2t} - \mathbf{D}\mathbf{y}_{1t} &= \mathbf{u}_{2t}. \end{aligned}$$

Here, $\mathbf{D} = \mathbf{D}(1)$ is an $n \times k$ matrix obtained by replacing L by 1 for some matrix polynomial $\mathbf{D}(\mathsf{L})$ in the lag operator. Moreover, $\mathbf{u}_t = (\mathbf{u}'_{1t}, \mathbf{u}'_{2t})' = \mathbf{F}(\mathsf{L})\boldsymbol{\varepsilon}_t$ for some other matrix polynomial $\mathbf{F}(\mathsf{L})$. Explicit definitions of $\mathbf{D}(\mathsf{L})$ and $\mathbf{F}(\mathsf{L})$ can be found in Watson (1994, p. 2875). Hence, $\{\mathbf{u}_t\}$ is a stationary moving-average process. In this representation, the elements $\mathbf{y}_{1t}$ are the common trends and $\mathbf{y}_{2t} - \mathbf{D}\mathbf{y}_{1t}$ are the I(0) stationary combinations of the data. The triangular representation has been particularly advocated by Phillips (1991), and a number of strong theoretical results have been obtained using it.

The estimation and testing theory of linear cointegrating systems is now well developed. It makes systematic use of functional limit results and expansions in terms of the multidimensional Wiener process akin to the introductory results for the scalar I(1) process considered in the preceding sub-section. We have found the review of Watson (1994) useful. It could also be mentioned that the cointegration theory has been extended to systems containing I(2) series; see Johansen (1992). Empirical aspects of linear cointegration are covered in Juselius (2006).

4. Nonlinear and nonstationary processes

This is meant to be the main focus of this survey. We start with the scalar case and consider two extensions of the unit root model in Sections 4.1 and 4.2. The nonlinear error correcting model is treated in Section 4.3 and a general nonlinear cointegrating regression model in Sections 4.4 and 4.5 with estimation in the parametric and in the nonparametric case, respectively.

4.1. Nonlinear I(1) processes

There are obvious difficulties in attempting to generalize the concepts discussed in the preceding section to nonlinearity because so many components are basically linear. The differencing operation is linear and so are consequently the definitions of I(1) and I(2). Furthermore, the cointegrating relationships are linear combinations of nonstationary variables. The various cointegrating systems including the error correction model

and the triangular representation are linear as well. It may be added that even though an I(1) process is invariant to linear transformations, it is not invariant to nonlinear transformations.

There are many "ad hoc" answers to the question of whether or not a nonlinear series can be I(1). A familiar example is to state that an interest rate series or any bounded time series cannot be I(1) because it is bounded below, whereas a random walk is unbounded. However, suppose that $\{\varepsilon_t\}$ is a positive series, $y_t = \sum_{s=0}^{t} \varepsilon_s$, then certainly $\{y_t\}$ is a random walk. It is I(1) because its difference is stationary, $\Delta y_t = \varepsilon_t$, but it is bounded below. The example may be thought to be objectionable because $\{y_t\}$ will have a trend in mean, as well as in variance, and all changes are positive. One could replace it by a more sophisticated model in which $\{y_t\}$ is a more standard random walk, with zero mean, some negative changes but reverts to a Markov process with a reflecting barrier as y_t gets small enough. Its properties will still be dominated by those of the random walk. Adding an upper boundary can be handled similarly. Nicolau (2002) has a discussion of bounded "near random walks."

Nevertheless, the concepts of I(1) and I(0) appropriate for nonlinear processes have to be rather more subtle. In a sense, if one wants to keep I(1) as an accumulation of I(0), the problem may be said to be not with I(1), but rather to find an appropriate nonlinear definition of I(0).

A difficulty is that even in the linear case, there is no concensus on how I(0) should be defined. Davidson (2009) lists five possible definitions, where the one mentioned in Section 3.1 is essentially one of them. Not requiring $\{y_t\}$ to be generated by a linear model, Davidson (2009) (see also Davidson (2002)) defines $\{y_t\}$ to be (possibly nonlinearly) I(0) if

$$\sigma_T^{-1} \sum_{t=1}^{[rT]} (y_t - \mathsf{E} y_t) \xrightarrow{d} w_r \tag{17}$$

where $\sigma_T^2 = \mathrm{var}(\sum_{t=1}^{T} y_t)$, and $\{w_r, 0 \le r \le 1\}$ denotes the Wiener process. Many of the linear definitions of I(0) are sufficient to guarantee this convergence, but Davidson prefers to state this directly as the defining property, thus loosening the bond to linearity. Irrespective of the definition of I(0) used, it seems to be difficult to construct consistent and asymptotically correctly sized tests for the I(0) hypothesis. In a sense, the problem is ill-posed (cf. Leeb and Pötscher, 2001; Pötscher, 2002). For a full discussion and a possible solution using simulation-based tests, we refer to Davidson (2009).

Another proposal of defining I(0) similar to the relationship (17) is concerned with α-mixing (Baghli, 2000; Escribano and Mira, 2002): A sequence $\{x_t\}$ is (possibly nonlinear) I(0) if it is an α-mixing sequence, whereas the series defined by $y_t = \sum_{s=1}^{t} x_s$ is not α-mixing. The series $\{y_t\}$ could then be said to be (possibly nonlinear) I(1).

If a process $\{x_t\}$ is α-mixing, then essentially

$$|\Pr\{x_t \in A_1, x_{t+\tau} \in A_2\} - \Pr\{x_t \in A_1\} \Pr\{x_{t+\tau} \in A_2\}| \to 0$$

as $\tau \uparrow \infty$ for all pairs of sets A_1, A_2, so $x_t, x_{t+\tau}$ becomes independent as τ gets large. If $\{x_t\}$ is not mixing, then it displays "some persistence of memory." Baghli (2000) states that the α-mixing property can be indirectly inferred by ensuring that a series verifies

Herrndorf's functional central limit theorem. This can be done using several standard tests including that based on the KPSS statistic (Kwiatkowski et al., 1992), see also Dufrénot and Mignon (2002).

In an interesting recent publication, Rico and Gonzalo (2010) forward the concept of summability: A stochastic process $\{y_t\}$ with positive variance is said to be summable of order δ if there exists a nonrandom sequence $\{m_t\}$ such that

$$S_T = \frac{1}{T^{\frac{1}{2}+\delta}} L(T) \sum_{t=1}^{T} (y_t - m_t) = O_p(1)$$

as $T \to \infty$, where δ is the minimum real number that makes S_T bounded in probability and $L(T)$ is a slowly varying function. They show that this concept generalize the concept of I(d) in the linear case and go on to establish the order of summability for a number of nonlinear models.

A rather different approach is taken in Karlsen and Tjøstheim (2001). They consider the I(1) processes rather than I(0). A generalized I(1) class containing both linear and nonlinear models is associated with the class of null recurrent Markov chains. The starting point is again the simple random walk (8). The two basic properties that Karlsen and Tjøstheim try to extend to a larger class of nonlinear I(1) type processes are (i) the persistence of the random walk (its nonstationarity); and (ii) the possibility of establishing central limit results such as the ones discussed in Davidson (2002) and Baghli (2000), but not necessarily with convergence to the Wiener process.

The random walk is a linear process and a Markov chain. The Markov chain property also holds for the nonlinear generalization

$$y_t = g(y_{t-1}) + \varepsilon_t \quad t \geq 1 \tag{18}$$

and such a process can be both stationary and nonstationary. If $|g(x)| \leq c\,|x|$ for some $c < 1$ when $|x|$ is large enough, then (Meyn and Tweedie, 1993) there exists an initial distribution for y_0, so that $\{y_t\}$ becomes stationary if started with this distribution, and property (i) above is not fulfilled; see Section 2.1 for more details. On the other hand, if g is such that $\{y_t\}$ is explosive, e.g., $g(x) = x^2$, then property (ii) cannot be satisfied in general; at least not in a nonparametric estimation context, because $\{y_t\}$ is then a transient Markov chain. A crucial property for $\{y_t\}$ to have for condition (ii) to hold is that it should be recurrent. This means that if $y_s = y$ for a certain time point s, then the Markov chain $\{y_t\}$ is guaranteed to be in an arbitrary small neighborhood around y with probability one at a future time point; the process recurs or regenerates. We refer to Karlsen and Tjøstheim (2001) for a more precise statement.

Under relatively weak regularity conditions, Karlsen and Tjøstheim derive a central limit theorem for sums of the type $\sum_{s=1}^{T} h(y_s)$ properly scaled, where h is a function satisfying some moment conditions. The key to this derivation is to use the recurrence property of the Markov chain to decompose the above sum as

$$\sum_{s=1}^{T} h(y_s) = \sum_{s=1}^{\tau_1} h(y_s) + \sum_{s=\tau_1+1}^{\tau_2} h(y_s) + \cdots + \sum_{s=\tau_n+1}^{T} h(y_s)$$

corresponding to the recurrence times $\tau_1, \tau_2, \ldots, \tau_n \leq T$; i.e., the time points of the regenerations of the chain. Clearly, $n \to \infty$ as $T \to \infty$, but at a slower rate. Due to the Markov property, the components $\sum_{s=\tau_i+1}^{\tau_{i+1}} h(y_s)$, $i = 1, \ldots, n$ are independent and identically distributed, and this can be used to prove a central limit result under the additional assumption that the distribution of the recurrence time intervals $S_i = \tau_i - \tau_{i-1}$ should not have too heavy tails. More specifically, $\Pr\{S_i > s\}$ is essentially of the order $s^{-\beta}$, $0 < \beta < 1$, so that $\mathsf{E}S_i^k < \infty$ for $k < \beta$. This property is named β-null recurrence in Karlsen and Tjøstheim (2001). The random walk corresponds to $\beta = 0.5$, as was established by Kallianpur and Robbins (1954).

Both parametric and nonparametric estimation can be handled by this technique, but Karlsen and Tjøstheim have restricted themselves to nonparametric estimation of a density function and the conditional mean function, covering the estimation of g in (18) as a special case. A very different approach based on random walk-like processes and local time of the Wiener process has been used by Park and Phillips (1999, 2001). Their method is outlined in more detail in Section 4.4. See also Xia (1998).

The class of recurrent Markov chains is subdivided into positive and null recurrent chains, depending on whether the expected recurrence time $\mathsf{E}S_i$ is finite or not. The positive recurrent case has $\mathsf{E}S_i < \infty$ ($\beta = 1$ in the above) and corresponds to stationarity, whereas the null recurrent case can be associated with a nonlinear extension of I(1). As already mentioned, the random walk is null recurrent with $\beta = 0.5$. A unit root AR(p) process can be cast as a p-dimensional Markov chain, and in Myklebust et al. (2011), it is shown that it is β-null recurrent with $\beta = 0.5$ under weak assumptions. This paper also contains a characterization of vector autoregressive time series as to when they are β-null recurrent, recurrent but not β recurrent, and transient. But the null recurrent class is not restricted to linear processes, and it has the useful invariance property that if $\{y_t\}$ is null recurrent (β-null recurrent) then the transformed process $\{h(y_t)\}$ is null recurrent (β-null recurrent) for an arbitrary one-to-one transformation h. Such an invariance property does not hold for the "ordinary" I(1) class of processes. The class of β-null recurrent processes satisfies both (i) and (ii) above, but this set-up is restricted by the fact that it must be possible to embed $\{y_t\}$ in a Markov chain framework, and only one unit root is allowed. One obvious class of examples, however, is obtained by considering $\{h(y_t)\}$, where h is one-to-one and $\{y_t\}$ is a simple random walk or an AR(p) unit root process. A threshold process whose far out behavior is a random walk would be another example of a null recurrent process and such threshold unit root processes are treated in Gao et al. (to appear), where they are also compared to other types of nonstationary threshold processes.

4.2. Stochastic unit root models

As an alternative to the nonlinear unit root type processes considered in the preceding sub-section, one can replace a unit root with a stochastic unit root and the term $1 - \mathsf{L}$ in (7) by $(1 - \rho_t \mathsf{L})$, where ρ_t is a stochastic process, but it is constrained to stay near one. This is a time-varying parameter model rather than being strictly nonlinear. Granger and Swanson (1997) and Leybourne et al. (1996) take ρ_t to be AR(1) with mean near one. Thus, occasionally the process will be stationary; in other periods, it will be somewhat explosive, but "on average" will be a unit root process. Standard tests, such as the augmented Dickey-Fuller test, cannot distinguish between an exact unit root and these

stochastic unit root (STUR) processes. Examples can be found where STUR models have superior forecasting properties over alternative univariate models.

In the AR(1) case, the stochastic unit root process can be written as

$$y_t = \phi_t y_{t-1} + \varepsilon_t.$$

with ϕ_t close to 1 in some sense or another. Using the specification given in Bec et al. (2008), they have been called autoregressive conditional root (or ACR) models. The authors consider, for instance, the simple form where $\phi_t = \rho^{s_t}$, ρ is a real number, s_t is a binary variable, taking values zero or one, and $\{\varepsilon_t\} \sim$ i.i.d.$(0, \sigma^2)$, possibly normally distributed. Moreover, $\{s_t\}$ will be a stochastic process, so that $\{y_t\}$ will be generated as a random walk if $s_t = 0$ and as a stationary AR(1) if $|\rho| < 1$ and $s_t = 1$. Conditions for stationarity are established and estimation is discussed. A related model has been analyzed by Gouriéroux and Robert (2006).

4.3. Nonlinear error-correction models

In Section 3.2, it was seen that there are two main representations of a linearly cointegrated system such that they could serve as a basis for nonlinear extensions. The error-correction model seems to be the one that has most often been used as a starting point, often with the nonlinear operation implemented only for the stationary process $\mathbf{x}_t$ in (15). We shall look at the nonlinear error correction (NLEC) model in this section. The NLEC model contains differences $\Delta \mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$ that can be considered a linear operation. In the next section, we look at the more general problem of establishing nonlinear relationships directly on I(1) type variables.

Most nonlinear extensions of the error correction model have been concerned with replacing the linear term $\mathbf{C}\mathbf{y}_{t-1}$ in (14) by a nonlinear generalization. However, Ripatti and Saikkonen (2001) consider a model where the intercept is contained in the cointegration space and is smoothly time-varying. This is obviously a form of nonlinearity, and Ripatti and Saikkonen (2001) use their model to test for a smoothly changing cointegration relationship.

We shall only consider the case of a bivariate process $\{\mathbf{y}_t\} = \{(y_{1t}, y_{2t})\}$ in (14). If $\{y_{1t}\}$ and $\{y_{2t}\}$ are both I(1), then they are linearly cointegrated if there is a constant vector $\boldsymbol{\alpha}$ such that $x_t = \boldsymbol{\alpha}' \mathbf{y}_t$ is I(0) (thinking of the linear I(0) class as in Section 3.1). It is generally true that if $\{x_t\}$ is stationary, then $\{g(x_t)\}$ is also stationary; assuming the mean and variance exist. A bivariate nonlinear error-correction (NLEC) model extending (15) takes the form

$$\Delta \mathbf{y}_t = \boldsymbol{\delta} g(x_{t-1}) + \sum_{i=1}^{p-1} \boldsymbol{\Psi}_i \Delta \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t$$

where $\boldsymbol{\delta} = (\delta_1, \delta_2)'$ is a two-dimensional vector and g is a function such that $g(0) = 0$ and $\mathsf{E}g(x_t)$ exists. The function g can be estimated nonparametrically or by assuming a particular parametric form. Escribano (1986, 2004) used a cubic function of x_t in a UK money demand equation and achieved a parsimonious model. Such a polynomial may

be viewed as a Taylor series approximation of

$$\beta_1 x_t + \beta_2 x_t (1 + \exp\{-\gamma(x_t - c)\})^{-1}$$

around $\gamma = 0$, which gives another form of NLEC.

An appealing form of NLEC models uses threshold error-corrections. This device was originally introduced by Balke and Fomby (1997), and it forms an important special case of the multivariate threshold model. Consider (5), but assume for simplicity that $\Gamma_{ij} = \mathbf{0}$ for all i and j. Assume that $r = 3$ and rewrite (5) as follows:

$$\Delta \mathbf{y}_t = \sum_{j=1}^{3} \left(\mu_j + \Pi_j \mathbf{y}_{t-1} + \sum_{k=1}^{p-1} \Psi_k^{(j)} \Delta \mathbf{y}_{t-k} + \varepsilon_{jt} \right) I(c_{j-1} < s_t \leq c_j)$$

for $p \geq 2$. When $p = 1$, the weighted sum of lagged first differences equals zero. Note that this threshold model is conceptually different from the one in Gao et al. (to appear). Assume that in each regime, $\mathbf{y}_t$ is integrated of order one ($\Delta \mathbf{y}_t$ stationary in the mean) and the $m \times m$ matrix Π_j can be written as $\Pi_j = \mathbf{A}_j \mathbf{B}'$, $j = 1, \ldots, r$, with $\operatorname{rank}(\mathbf{A}_j) < m$, and that s_t is continuous and stationary as before. If $m = \dim(\mathbf{y}_t) = 2$, say, then $\mathbf{A}_j = (\alpha_{1j}, \alpha_{2j})'$ and $\mathbf{B} = (1, \beta_2)$, and the variables y_{1t} and y_{2t} are cointegrated with cointegrating vector $\mathbf{B}$. This model is called the threshold vector error correction (TVEC) model. The strength of the attraction varies in the three regimes according to $\mathbf{A}_j$. An interesting case in applications is the one in which $r = 3$ and $\mathbf{A}_2 = \mathbf{0}$. This means that the model has three regimes and cointegration is present at high and low values of the threshold variable s_t but not in the middle. At the same time, the intercept is restricted in the cointegrating relationship, so that

$$\Delta \mathbf{y}_t = \sum_{j=1}^{3} \left\{ \mathbf{A}_j (\mathbf{B}' \mathbf{y}_{t-1} - \mu_j) + \sum_{k=1}^{p-1} \Psi_k^{(j)} \Delta \mathbf{y}_{t-k} + \varepsilon_{jt} \right\} I(c_{j-1} < s_t \leq c_j).$$

The simplest model of this sort is the bivariate band-TVEC model. It has a band around the line $x = 0$, in which there is no cointegration, then upper and lower bands, in which cointegration occurs, although possibly with different "strengths."

The threshold error-correction model has been further developed by several authors. Hansen and Seo (2002) provided a testing theory for the case where the cointegrating vector is estimated, and they treat a general multivariate case. Saikkonen (2005) derived stability results for the general NLEC.

More general switching mechanisms than the threshold one have been treated in Bec and Rahbek (2004). Finally, there are a number of applications of NLEC models, e.g., Bec et al. (2004), Baghli (2004), and Escribano (2004). There are still a number of unsolved problems of statistical inference in these models.

4.4. Parametric nonlinear regression with a nonstationary regressor

For a stationary regressor x_t, the parametric nonlinear regression model

$$y_t = g(x_t, \boldsymbol{\theta}) + u_t \tag{19}$$

where g is known, $\boldsymbol{\theta}$ is an unknown parameter vector and u_t is stationary, can be analyzed using fairly standard methods. This is not the case in the nonlinear and nonstationary cases.

For the linear nonstationary regression case, it was seen in Section 3.1 that the asymptotics of sums of type $\sum y_t^2$ and $\sum \varepsilon_t y_{t-1}$ need to be evaluated when $\{y_t\}$ is an I(1) process. Properly scaled these sums converge to integrals over a Wiener process.

In this sub-section, which is based on Park and Phillips (1999, 2001), we shall consider the rather general regression model (19), in which $\{u_t\}$ is a martingale increment process and $\{x_t\}$ an integrated process such that $\Delta x_t = v_t$. Here, $\{v_t\}$ could be a moving-average process or more generally a process such that

$$v_T(r) = \frac{1}{\sqrt{T}} \sum_{s=1}^{[rT]} v_s \tag{20}$$

converges to a Wiener process w_r. Again, $[rT]$ is the integer part of rT. Moreover, it is assumed that

$$(u_T(r), v_T(r)) \xrightarrow{d} (w_{1r}, w_{2r}) \tag{21}$$

where $\{(w_{1r}, w_{2r})\}$ is a vector Wiener process, and

$$u_T(r) = \frac{1}{\sqrt{T}} \sum_{s=1}^{[rT]} u_s.$$

It should be noted that this set-up with $\{x_t\}$ being an I(1) type process excludes the possibility of analyzing the model

$$y_t = g(y_{t-1}, \boldsymbol{\theta}) + u_t \tag{22}$$

because the class of I(1) processes is not invariant under a general nonlinear transformation g, and because $\{y_t\}$ enters on both sides of the equality (22), it cannot be of I(1) type. Nonlinear and nonstationary AR models of type (22) will be estimated nonparametrically in the next sub-section.

We will consider a least squares estimator $\widehat{\boldsymbol{\theta}}_T$ of $\boldsymbol{\theta}$ in (19); that is, $\widehat{\boldsymbol{\theta}}_T$ is taken to minimize

$$Q_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} \{y_t - g(x_t, \boldsymbol{\theta})\}^2. \tag{23}$$

Let $\dot{Q}_T = \partial Q_T / \partial \boldsymbol{\theta}$ and $\ddot{Q}_T = \partial Q_T / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$. The asymptotic analysis of $\widehat{\boldsymbol{\theta}}_t$ takes as its starting point the Taylor expansion

$$\dot{Q}_T(\widehat{\boldsymbol{\theta}}_T) = \dot{Q}_T(\boldsymbol{\theta}_0) + \ddot{Q}_T(\boldsymbol{\theta}_T)(\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)$$

where $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_T$ is an intermediate value determined by the mean value theorem. Using a scaling factor ν_T and the fact that $\dot{Q}(\widehat{\boldsymbol{\theta}}_T) = 0$, this implies

$$\nu_T(\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) = [\nu_T^{-1}\ddot{Q}_T(\boldsymbol{\theta}_0)\nu_T^{-1}]^{-1}\nu_T\dot{Q}_T(\boldsymbol{\theta}_0) + o_p(1).$$

It is seen that this leads to the evaluation of sums of type $\sum_t h_1(x_t, \boldsymbol{\theta}_0)$ and $\sum_t h_2(x_t, u_t, \boldsymbol{\theta}_0)$ for some functions h_1 and h_2 depending on the function g and its derivative. The evaluation of such sums is a crucial part of the analysis, and many of its aspects are covered in Park and Phillips (1999). The heuristics of Section 3.1 can be used for homogeneous functions with the property that $h(\lambda x) = \lambda^k h(x)$ for a scalar λ and some k, for example $h(x) = x^k$. Then,

$$\frac{1}{T^{1+k/2}}\sum_{t=1}^{T} x_t^k = \sum_{t=1}^{T}\Big(\frac{x_t}{\sqrt{T}}\Big)^k \cdot \frac{1}{T} \xrightarrow{d} \int_0^1 w_{2r}^k dr. \tag{24}$$

Here, $\{w_{2r}\}$ is the Wiener process appearing in (21) and $\{x_t\}$ plays the role of $\{y_t\}$ in (10) and (11). In general, it can be shown (Park and Phillips, 1999) that for so-called regular functions (including continuous and piecewise continuous functions)

$$\frac{1}{T}\sum_t h\Big(\frac{x_t}{\sqrt{T}}\Big) \xrightarrow{d} \int_0^1 h(w_{2r})dr.$$

Park and Phillips (1999, 2001) consider altogether four classes of functions:

1. Integrable functions with the property that $\int_{-\infty}^{\infty} h(x)dx$ exists and is finite, so that $h(x) \to 0$ at a fast enough rate as $x \to \pm\infty$.
2. Asymptotic homogeneous functions, having the property

$$h(\lambda x) = k(\lambda)H(x) + R(x, \lambda) \tag{25}$$

   where $H(x)$ is a homogeneous function so that $H(\lambda x) = k(\lambda)H(x)$ for some $k(\lambda)$, $R(x, \lambda)$ is dominated by $H(x)$ when $|x|$ gets large.
3. Asymptotic exponential functions, where h grows to infinity with the speed of an exponential function.
4. Super exponential functions, where h grows to infinity faster than the simple exponential.

These classes of functions lead to rather different types of behavior for $\sum_t h(x_t)$, and unlike the linear and homogeneous case, integrals of functions of a Wiener process do not suffice. One needs to introduce the concept of local time of the Wiener process,

$$L(t, s) = \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon}\int_0^t I(|w_r - s| < \varepsilon)dr$$

where I is the indicator function. It may be noted that $L(t,s)$ is a random process in both t and s. It essentially measures the time that w_r spends close to s in the time interval $[0,t]$. It can be introduced in a much more general setting than the Wiener process. It can be made meaningful both for Markov processes and semimartingales. Much of its importance stems from the so-called occupation time formula. It states that if h is locally integrable, then

$$\int_0^t h(w_r)dr = \int_{-\infty}^{\infty} h(s)L(t,s)ds. \tag{26}$$

which again is valid in a much more general setting, see for instance Revuz and Yor (1994).

It is interesting to note that the move from the analysis of stationary series to unit root processes required the introduction of mathematical techniques based on the Wiener process. Now moving further to the analysis of nonlinear transformations of unit root processes involves further new mathematics, involving local time.

Park and Phillips (1999) prove under some regularity conditions that if h is integrable and x_t is an integrated process such that $\Delta x_t = v_t$ with $\{v_t\}$ as in (20), then

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} h(x_t) \xrightarrow{d} \left(\int_{-\infty}^{\infty} h(s)ds\right) L(1,0) \tag{27}$$

as $T \to \infty$.

This result means that $\sum_t h(x_t)$ spreads out at a rate of (is balanced by a scaling factor) $\sqrt{T}$. Moreover, the integrability of h implies that h tends to zero far out, and only observations at zero are exploited in the accompanying Wiener process as indicated by the local time variable $L(1,0)$. The behavior is very different in the homogeneous case. Indeed, again under some regularity conditions, if h is asymptotically homogeneous satisfying the decomposition (25), then

$$\frac{1}{Tk(\sqrt{T})}\sum_{t=1}^{T} h(X_t) \xrightarrow{d} \int_0^1 H(w_r)dr = \int_{-\infty}^{\infty} H(s)L(1,s)ds \tag{28}$$

as $T \to \infty$. The last equality follows from the occupation time formula (26). Moreover, (24) is a special case of (28) with $R(x,\lambda) = 0$, $h(x) = H(x) = x^k$ and $k(\lambda) = \lambda^k$ in (25).

Park and Phillips (1999) also derive a theorem for the exponential case which again leads to radically different behavior. We refer to their paper for details.

For an integrable $g(x,\theta)$ with a scalar parameter θ in (19), results such as in (27), under regularity conditions stated in Park and Phillips (2001), lead to the following

central limit theorem for $\widehat{\theta}_T$ minimizing (23):

$$T^{1/4}(\widehat{\theta}_T - \theta_0) \xrightarrow{d} \left(L(1,0) \int_{\infty}^{\infty} \dot{g}(s,\theta_0)^2 ds \right)^{-1/2} w_1$$

where $\dot{g} = \partial g/\partial\theta$, and where w_1, the Wiener process at time 1, is a standard normal random variable. The convergence rate is seen to be slower than the standard parametric convergence rate $T^{-1/2}$ of the stationary case. It comes from the scaling factor of $T^{1/2}$ in (27) and from a corresponding scaling factor $T^{1/4}$ for sums of type $\sum h(x_t)u_t$, with u_t defined in (19), for an integrable function h.

The analogous result for a homogeneous type $g(x,\boldsymbol{\theta})$ with a vector parameter $\boldsymbol{\theta}$ in (19) is given by (under a number of regularity conditions)

$$\sqrt{T}\dot{k}(\sqrt{T})'(\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} \left(\int_0^1 \dot{H}(w_{2r},\boldsymbol{\theta}_0)\dot{H}(w_{2r},\boldsymbol{\theta}_0)' dr \right)^{-1} \int_0^1 \dot{H}(w_{2r},\boldsymbol{\theta}_0) dw_{1r}. \quad (29)$$

Here, H is the homogeneous part of $g(x,\boldsymbol{\theta})$ defined analogously to (25), $\dot{H} = \partial H/\partial\boldsymbol{\theta}$, k is the asymptotic order of $g(x,\boldsymbol{\theta})$ as in (25) (it may depend on $\boldsymbol{\theta}$) and $\dot{k}$ is defined as the corresponding asymptotic order of $\dot{g}(x,\boldsymbol{\theta})$, so that $\dot{g}(\lambda x,\boldsymbol{\theta}) = \dot{k}(\lambda)\dot{H}(x,\boldsymbol{\theta})$ asymptotically as x gets large. Finally, (w_{1r}, w_{2r}) is the pair of Wiener processes appearing in (21). In the scalar linear case $g(x,\theta) = \theta x$, $\dot{g}(x,\theta) = x$, such that $k(\lambda) = \dot{k}(\lambda) = \lambda$. This gives $\sqrt{T}\dot{k}(\sqrt{T}) = \sqrt{T}\sqrt{T} = T$ leading to the convergence rate of T^{-1} as in Section 3.1, faster than the standard stationary rate. It is also easy to check that the formula in (29) reduces to the standard formula in the linear case.

There are a host of challenging problems in this field. For example, even though the regression relationship (19) is sometimes called a nonlinear cointegrating relationship, it does not really have the same symmetry in y and x as in the linear cointegrating case. For that purpose transformations of both y_t and x_t may be required. Granger and Hallman (1991) have considered this problem A very recent contribution is Goldstein and Stigum (to appear) considering joint transformations of (y_t, x_t). Extending the analysis beyond the bivariate case forms another difficulty. Such an extension is far from trivial.

Saikkonen and Choi (2004) and Choi and Saikkonen (2004) consider estimation and testing of the model (19), where g is a smooth transition function. They apply another type of asymptotics, so-called triangular array asymptotics (cf. Andrews and McDermott, 1995). In this kind of asymptotics, the actual sample size is fixed at T_0, say, and the model is embedded in a sequence of models depending on a sample size T which tends to infinity. The embedding is obtained by replacing the I(1) regressor x_t in (19) by $(T_0/T)^{1/2}x_t$. This change leads to a central limit theorem for the least squares estimate $\widehat{\boldsymbol{\theta}}$ with rate $T^{-1/2}$ under some regularity conditions including a three times differentiability condition on the function g in (19). It is seen that the triangular array asymptotics is rather different from that used in Park and Phillips (1999, 2001).

4.5. Nonparametric estimation in a nonlinear cointegration type framework

Karlsen et al. (2007, 2010) consider nonparametric estimation in a nonlinear nonstationary environment which in some respects is wider than that of Park and Phillips but in other respects more narrow. The class of models is defined by

$$y_t = g(x_t) + u_t \tag{30}$$

where x_t is nonstationary and β-null recurrent as defined in Section 4.1, u_t is a stationary infinite-order moving-average process or a Markov chain. In contrast to the set-up in Section 4.4, one can now allow $x_t = y_{t-1}$. Karlsen and Tjøstheim (2001) in fact discuss estimation in this case. The function g is unknown, and the task is to estimate it nonparametrically. Except for trivial choices of g (e.g., $g =$ constant), the process $\{y_t\}$ will be nonstationary, but it will not be β-null recurrent, as it is not even a Markov chain. The analysis in Karlsen et al. (2007) is carried out in two cases: the case in which $\{x_t\}$ and $\{u_t\}$ are independent, and the one in which dependence is allowed between them. At its present state, the dependence modelling also requires a boundedness condition for $\{u_t\}$.

The function $g(x)$ in (30) is estimated nonparametrically using the Nadaraya-Watson estimator

$$\widehat{g}(x) = \sum_{t=1}^{T} y_t K_h(x_t - x) / \sum_{t=1}^{T} K_h(x_t - x)$$

where $K_h(u) = h^{-1}K(h^{-1}u)$ is the kernel with bandwidth h. Karlsen et al. (2007) prove that

$$\left\{h \sum_{t=1}^{T} K_h(x_t - x)\right\}^{1/2} \left\{\hat{g}(x) - g(x) - \text{ bias term}\right\} \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \int K^2(s)ds\right) \tag{31}$$

as $T \to \infty$. Here, $\sigma^2 = \text{var}(u_t)$. The bias term tends to zero as $T \to \infty$, and it is explicitly given in Karlsen et al. (2007). The convergence of $\widehat{g}(x)$ to $g(x)$ is slower than in the stationary case. This is easy to explain since the null recurrence of $\{x_t\}$ means that it takes more time for the process to return to a neighborhood of the point x, and it is the points in the neighborhood of x which are used in the nonparametric estimation. Roughly speaking, the sample size is in effect reduced from T to T^β with $\beta = 1/2$ if $\{x_t\}$ is a random walk. Then, the rate of convergence for $\widehat{g}(x)$ equals $T^{-1/4}h^{-1/2}$. For a fixed h, this is seen to be the same rate as the parametric estimation rate of $\widehat{\boldsymbol{\theta}}$ with an integrable function $g(x, \boldsymbol{\theta})$ in (19). The kernel function K plays the role of the integrable function in the nonparametric case. It should also be noted that in Karlsen et al. (2007), the so-called Mittag-Leffler process is the analogy of the local time process $L(t, 0)$. The relationship between these processes needs to be further explored in the Markov case. A proper understanding of it would lead to more

general and unified procedures. Wang and Phillips (2009a,b) use the local time as an alternative to obtain an asymptotic theory of the nonparametric estimates treated in Karlsen et al. (2007), but again this approach does not allow x_t to be replaced by y_{t-1} in (30). Finally, it should be noted that in contrast to the majority of limit theorems in this chapter the limit in (31) is Gaussian. This is due to the stochastic scaling used.

In Karlsen et al. (2007), and Wang and Phillips (2009a,b), a fixed bandwidth $h = h_T$ independent of x in (31) has been used in the estimation. Since in the nonstationary case, the data points are widely scattered, using a variable bandwidth could be advantageous. Actually, the early paper by Yakowitz (1993) uses nearest neighbor estimation. A follow-up to this is Sancetta (2009), who attacks quite general conditional nonparametric problems using nearest neighbors. Both papers are limited to proving consistency results. The recent paper by Bandi et al. (2011) considers automated bandwidth choice in a nonstationary kernel estimation context.

An attempt to establish a theory for specification testing is contained in Gao et al. (2009a) for the time series regression case and in Gao et al. (2009b) for the time series autoregressive case. They consider the nonlinear AR model

$$x_t = g(x_{t-1}) + \varepsilon_t \tag{32}$$

and test the null hypothesis $g(x) = x$, a linear unit root process, against a stationary alternative $g(x) = x + g_1(x)$ for some $g_1(x)$. They give nonlinear examples in which their test has better power than the standard Dickey-Fuller test. Moreover, in a nonlinear cointegration type situation with

$$y_t = g(x_t) + u_t \tag{33}$$

they test whether the function $g(x)$ equals $g_1(x, \boldsymbol{\theta})$, a known parametric function. This includes the case where $g_1(x, \boldsymbol{\theta})$ is linear.

Rather restrictive assumptions of Gaussianity and i.i.d. are used for the error processes $\{u_t\}$ and $\{\varepsilon_t\}$ in (32) and (33). They are also assumed to be independent processes. In the time series regression case, a more general model and weakened assumptions are considered in Wang and Phillips (2010) using local time arguments. As already mentioned, Choi and Saikkonen (2004) have considered a parametric cointeg ration test for linearity using the triangular array asymptotics.

5. Time-varying parameters and state-space models

5.1. Introduction

A very general model is

$$y_t = g(\mathbf{z}_t, \boldsymbol{\theta}, \varepsilon_t)$$

where g is a known function, $\mathbf{z}_t$ is a vector of explanatory variables possibly including lagged values of y_t, $\boldsymbol{\theta}$ is an unknown parameter vector, and ε_t is an error term. The

parameter $\boldsymbol{\theta}$ contributes to describing the dynamics of the process. In this sense, for a fixed $\boldsymbol{\theta}$, it represents the state of the system. Different values of $\boldsymbol{\theta}$, different states, may lead to quite different dynamics.

Sometimes, it is assumed that $\boldsymbol{\theta} = \boldsymbol{\theta}_t$ varies in time, yielding a time-varying dynamics, and then it becomes important to estimate $\boldsymbol{\theta}_t$ as a function of time, not least if one's objective is to make forecasts of future values of the series $\{y_t\}$. Time dependence can be introduced in two ways, deterministic or stochastic. The first option leads to nonstationarity of $\{y_t\}$. Unless relatively strict regularity conditions are imposed on the time-variation of $\{\boldsymbol{\theta}_t\}$, it is difficult to analyze in practice and to make forecasts. For instance, one may assume that $\{\boldsymbol{\theta}_t\}$ is constant in time except for sudden changes or breaks at (usually) unknown time points, or there may be a smooth parameterized transition, or $\{\boldsymbol{\theta}_t\}$ may be slowly time-varying in some specified way, for instance resulting in a slowly time-varying spectrum; see Priestley (1965), Dahlhaus (1997, 2001), Dahlhaus et al. (1999), and Dahlhaus and Subba Rao (2006).

In spite of the progress in this area, the other alternative of modelling $\{\boldsymbol{\theta}_t\}$ as a random process has been more common. Letting $\{\boldsymbol{\theta}_t\}$ be stochastic still allows $\{y_t\}$ to be stationary under some regularity conditions. Moreover, employing the structural properties of $\{\boldsymbol{\theta}_t\}$, if it is estimated, it can be predicted, which in turn can be used to make forecasts for $\{y_t\}$. Using such a set-up leads to so-called state-space processes, see e.g., Durbin and Koopman (2001) for an introduction. State-space models are sometimes divided into observation and parameter driven models using the terminology of Cox (1981); see also Davis et al. (2003, 2005). In observation driven models, the process $\{\boldsymbol{\theta}_t\}$ is generated by observations. One example is the conditional variance process $\{h_t\}$ in a GARCH model, where $\{h_t\}$ is an unobserved component process driven by $\{y_t\}$, although GARCH models are not usually thought of as being state-space models. In a parameter driven model, the observations are not involved in the driving mechanism for $\{\boldsymbol{\theta}_t\}$. The class of stochastic volatility models (Shephard, 2005) would be a typical example.

The case where the state space for $\{\boldsymbol{\theta}_t\}$ is continuous corresponds to smooth changes of the dynamics for the $\{y_t\}$-process. There is also a growing literature for the situation where the state space of $\{\boldsymbol{\theta}_t\}$ is discrete, and then usually finite. These models are usually called finite regime models or hidden Markov chain models, when a Markov assumption is added.

5.2. Nonlinear state-space models

Although there seems to be no consensus precisely as to what constitutes a nonlinear state-space model, several authors have considered models of the form

$$\mathbf{y_t} = \mathbf{a}(\boldsymbol{\theta}_t) + \mathbf{b}(\mathbf{z}_t) + \boldsymbol{\varepsilon}_t \tag{34}$$

$$\boldsymbol{\theta}_t = \mathbf{c}(\boldsymbol{\theta}_{t-1}) + \boldsymbol{\eta}_t \tag{35}$$

where $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are vector functions. A model that combines the $\mathbf{z}$-dependence and the time-varying parameter aspect is the one in which the above observational Eq. (34) is replaced by

$$\mathbf{y}_t = \mathbf{g}(\mathbf{z}_t, \boldsymbol{\theta}_t) + \boldsymbol{\varepsilon}_t \tag{36}$$

which, unlike (34), is nonadditive in $\mathbf{z}_t$ and $\boldsymbol{\theta}_t$. To our knowledge, such models have not been much treated in the literature. A related example, though, is the observation driven STAR model with a stochastically varying parameter, which has been analyzed in Anderson and Low (2006).

The conventional nonlinear state-space model (34)–(35) has been treated essentially by three approaches, the extended Kalman filter, the Kitagawa grid approximation, and Monte Carlo methods. These approaches complement each other in that different model assumptions are needed. A fourth method is based on Gaussian approximation using linearization as in the extended Kalman filter combined with Monte Carlo techniques.

The idea of the extended Kalman filter is to linearize $\mathbf{a}(\boldsymbol{\theta}_t)$ and $\mathbf{c}(\boldsymbol{\theta}_t)$ around $\boldsymbol{\theta}_{t|t-1}$ and $\boldsymbol{\theta}_{t|t}$, respectively. Here $\boldsymbol{\theta}_{t|s} = \mathbb{E}\{\boldsymbol{\theta}_t|\mathcal{F}_s^z \vee \mathcal{F}_s^y\}$, where $\mathcal{F}_s^z$ and $\mathcal{F}_s^y$ are the σ-algebras generated by $\{\mathbf{z}_u, u \le s\}$ and $\{\mathbf{y}_u, u \le s\}$, respectively. We then have

$$\mathbf{a}(\boldsymbol{\theta}_t) = \mathbf{a}(\boldsymbol{\theta}_{t|t-1}) + \frac{d\mathbf{a}}{d\boldsymbol{\theta}}(\boldsymbol{\theta}_{t|t-1})(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t|t-1})$$

where higher-order terms are neglected. Similarly,

$$\mathbf{c}(\boldsymbol{\theta}_t) = \mathbf{c}(\boldsymbol{\theta}_{t|t}) + \frac{d\mathbf{c}}{d\boldsymbol{\theta}}(\boldsymbol{\theta}_{t|t})(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t|t}).$$

Inserting these in (34) and (35), we have

$$\mathbf{y}_t = \mathbf{a}(\boldsymbol{\theta}_{t|t-1}) + \frac{d\mathbf{a}}{d\boldsymbol{\theta}}(\boldsymbol{\theta}_{t|t-1})(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t|t-1}) + \mathbf{b}(\mathbf{z_t}) + \varepsilon_t \tag{37}$$

and

$$\boldsymbol{\theta}_{t+1} = \mathbf{c}(\boldsymbol{\theta}_{t|t}) + \frac{d\mathbf{c}}{d\boldsymbol{\theta}}(\boldsymbol{\theta}_{t|t})(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t|t}) + \boldsymbol{\eta}_{t+1}. \tag{38}$$

Now, using the definitions of $\boldsymbol{\theta}_{t|t-1}$ and $\boldsymbol{\theta}_{t|t}$, these are functions of variables observed at time $t-1$ and t, and it is seen that the above equations can be identified with a linear time-varying parameter Kalman system with a time-varying intercept in the state equation, and a Kalman algorithm can then be set up.

If there is strong nonlinearity in the series, the first-order extended Kalman filter will not work too well. A useful alternative is the Kitagawa (1987) grid approximation. This method has the extra advantage that no Gaussian assumption is needed for the generating processes $\{\boldsymbol{\varepsilon}_t\}$ and $\{\boldsymbol{\eta}_t\}$. Such an assumption is crucial for the extended Kalman algorithm, since its derivation is still based on the simple formulas for conditional Gaussian distributions.

In the absence of a Gaussian assumption, the first two conditional moments no longer describe the contemporaneous conditional structure of the model. Instead of updating the first two conditional moments, the task is to update the entire density function $f(\boldsymbol{\theta}_t|\mathcal{F}_{t-1}^y)$ to $f(\boldsymbol{\theta}_{t+1}|\mathcal{F}_t^y)$. This is too ambitious, but, following Kitagawa, the update will be made on a finite grid of points $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(N)}$. This implies that the input consists of the $N+1$ values $f(\boldsymbol{\theta}_t = \boldsymbol{\theta}^{(i)}|\mathcal{F}_{t-1}^y)$, $i = 0, \ldots, N$, and the problem consists

in producing the update $f(\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}^{(i)}|\mathcal{F}_t^y)$, $i = 0, \ldots, N$. Using the Kitagawa grid approach, one can update the conditional density $f(\boldsymbol{\theta}_t|\mathcal{F}_{t-1}^y)$ to $f(\boldsymbol{\theta}_{t+1}|\mathcal{F}_t^y)$ for a finite (and fixed) set of grid points $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(N)}$ using numerical integration. For complex (and multivariate) problems, one may encounter numerical instabilities. A standard way of evaluating integrals is via Monte Carlo simulation. It is, therefore, perhaps not so surprising that recently a number of Monte Carlo methods have been developed where numerical integration is avoided in updating the filter density $f(\boldsymbol{\theta}_t|\mathcal{F}_{t-1}^y)$. Typically, the filter is updated for a set of stochastic values of $\boldsymbol{\theta}_t$ (as opposed to a fixed set of values in the Kitagawa method). This is often combined with importance sampling techniques to obtain the so-called particle filters.

In Koopman et al. (2005), Koopman and Ooms (2006), and Menkveld et al. (2007), the authors look at a nonlinear Gaussian system and use a Gaussian approximation to obtain computationally efficient algorithms, which they then have applied on a number of economic problems

All of the above is concerned with a continuous state space. There is also a large branch of literature, see e.g., Cappé et al. (2005), on the discrete case, again distinguishing between observation and parameter driven models. The class of hidden Markov chains belong to the latter category. The hidden Markov model represents a mixture of regimes in that different AR or other parametric models result for each value of $\boldsymbol{\theta}_t$. Such a mixture can be obtained by various means and can be extended to a mixture of other types of models. The modeling is typically made directly on the conditional density of $\{y_t\}$ given past values of $\{y_t\}$ and possible explanatory variables. This kind of models are often called mixture models (cf. Wong and Li, 2000). Estimation of parameters in nonlinear state-space models is still in its infancy, and the nonstationary case has hardly been touched upon. Maximum likelihood methods combined with importance sampling have been considered by Shephard and Pitt (1997) and by Durbin and Koopman (2000) and by Davis and Rodrigues-Yam (2005). Work on the central limit theorem, asymptotic distributions has been done by Bickel et al. (1998), Jensen and Petersen (1999), and Douc et al. (2004). A full overview of the statistical inference in hidden Markov chains can be found in Cappé et al. (2005).

References

Anderson, H.M., Low, C.N., 2006. Random walk smooth transition autoregressive models. In: Milas, C., Rothman, P., van Dijk, D. (Eds.), Nonlinear Time Series Analysis of Business Cycles. Elsevier, Amsterdam, pp. 247–281.

Andrews, D.W.K., McDermott, C.J., 1995. Nonlinear econometric models with deterministically trending variables. Rev. Econ. Stud. 62, 343–360.

Aue, A., Berkes, I., Horváth, L., 2006. Strong approximation for the sum of squares of augmented GARCH sequences. Bernoulli 12, 583–608.

Baghli, M., 2000. Modeling the FF/DM rate by threshold cointegration analysis. Stat. Inf. Stoch. Processes 3, 113–128.

Baghli, M., 2004. Modelling the FF/MM rate by threshold cointegration analysis. Appl. Econom. 36, 533–548.

Balke, N.S., Fomby, T.B., 1997. Threshold cointegration. Int. Econ. Rev. 38, 627–645.

Bandi, F., Corradi, V., Wilhelm, D., 2011. Nonparametric Nonstationary Autoregression and Nonparametric Cointegrating Regression: Automated Bandwidth Selection, Manuscript, Department of Economics, Johns Hopkins University.

Bec, F., Ben Salem, M., Carrasco, M., 2004. Tests for unit roots versus threshold specification with an application to the PPP. J. Bus. Econ. Stat. 22, 382–395.

Bec, F., Rahbek, A., 2004. Vector equilibrium correction models with nonlinear discontinuous adjustments. Econom. J. 7, 628–651.

Bec, F., Rahbek, A., Shephard, N., 2008. The ACR model: A multivariate dynamic mixture autoregression. Oxf. Bull. Econ. Stat. 70, 583–618.

Bickel, P., Bühlmann, P., 2003. What is a linear process? Proc. Natl. Acad. Sci. 93, 12128–12131.

Bickel, P.J., Ritov, A., Rydén, T., 1998. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. Ann. Stat. 26, 1614–1635.

Breidt, F.J., Davis, R.A., Hsu, N.-J., Rosenblatt, M., 2006. Pile-up probabilities for the Laplace likelihood estimator of a non-invertible first order moving average. In: Ho, H.-C., Lai, T.L., (Eds.), Memory of Ching-Zong Wei, IMS Lecture Notes, pp. 1–19.

Cappé, O., Rydén, T., Moulines, E., 2005. Inference in Hidden Markov Chains, Springer, New York.

Carrasco, M., Chen, X., 2002. Mixing and moment properties of various GARCH and stochastic volatility models. Econ. Theory 18, 17–39.

Chan, K.S., 1993. Consistency and limiting distribution of a least squares estimator of a threshold autoregressive model. Ann. Stat. 21, 520–533.

Choi, I., Saikkonen, P., 2004. Tests of linearity in cointegrating smooth transition regressions. Econ. J. 7, 341–365.

Cox, D.R., 1981. Statistical analysis of time series: Some recent developments. Scand. J. Stat. 8, 93–115.

Dahlhaus, R., 1997. Fitting time series models to nonstationary processes. Ann. Stat. 25, 1–37.

Dahlhaus, R., 2001. A likelihood approximation for locally stationary processes. Ann. Stat. 28, 1762–1794.

Dahlhaus, R., Neumann, M.H., von Sachs, R., 1999. Nonlinear wavelet estimation in time-varying autoregressive processes. Bernoulli 5, 873–906.

Dahlhaus, R., Subba Rao, S., 2006. Statistical inference for time-varying ARCH processes. Ann. Stat. 34, 1075–1114.

Davidson, J., 2002. Establishing conditions for the functional central limit theorem in nonlinear and semiparametric time series processes. J. Econ. 106, 243–269.

Davidson, J., 2009. When is a time series I(0)? In: Castle, J., Shephard, N., (Eds.), A Festschrift for David Hendry. Oxford University Press, Oxford, pp. 322–242.

Davis,R.A., Dunsmuir, W.T.M.,Streett, S.B., 2003. Observation-driven models for poisson counts. Biometrika 90, 777–790.

Davis, R.A., Dunsmuir, W.T.M., Streett, S.B., 2005. Maximum likelihood estimation for an observation driven model for poisson counts. Methodol. Comput. Appl. Probab. 7, 149–159.

Davis, R., Rodrigues-Yam, G., 2005. Estimation for state-space models; an approximate likelihood approach. Stat. Sin. 15, 381–406.

de Gooijer, J.G., 1998. On threshold moving-average models. J. Time Ser. Anal. 19, 1–18.

Douc, R., Moulines, E., Rydén, T., 2004. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. Ann. Stat. 32, 2254–3004.

Dufrénot, G., Mignon, V., 2002. Recent Developments in Nonlinear Cointegration with Applications to Macroeconomics and Finance, Kluwer, Amsterdam.

Durbin, J., Koopman, S.J., 2000. Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion). J. R. Stat. Soc. Ser. B 62, 3–56.

Durbin, J., Koopman, S.J., 2001. Time Series Analysis by State Space Methods, Oxford University Press, Oxford.

Engle, R.F., Granger, C.W.J., 1987. Cointegration and error correction: Representation, estimation and testing. Econometrica 55, 251–276.

Escribano, A., 1986. Identification and modeling of economic relationships in a growing economy, PhD Thesis, University of California, San Diego.

Escribano, A., 2004. Nonlinear error correction: the case of money demand in the United Kingdom (1878–2000). Macroecon. Dyn. 8, 76–116.

Escribano, A., Mira, S., 2002. Nonlinear error correction models. J. Time Ser. Anal. 23, 509–522.
Fan, J., Yao, Q., 2003. Nonlinear Time Series. Nonparametric and Parametric Methods, Springer, New York.
Francq, C., Zakoïan, J.-M., 2006. Mixing properties of a general class of GARCH(1,1) models without moment assumptions. Econ. Theory 22, 815–834.
Gao, J., King, M., Lu, Z., Tjøstheim, D., 2009a. Nonparametric specification testing for nonlinear time series with nonstationarity. Econ. Theory 25, 1869–1892.
Gao, J., King, M., Lu, Z., Tjøstheim, D., 2009b. Specification testing in nonlinear and nonstationary time series autoregression. Ann. Stat. 37, 3893–3928.
Gao, J., Tjøstheim, D., Yin, J., to appear. Estimation in threshold autoregressive models with a stationary and a unit root regime. J. Econ.
Goldstein, H., Stigum, B.P., to appear. Nonlinear cointegration in foreign exchange markets. J. Econ.
Gouriéroux, C., Robert, C.Y., 2006. Stochastic unit root models. Econ. Theory 26, 1052–1090.
Granger, C.W.J., 1981. Some properties of time series data and their use in econometric model specification. J. Econ. 16, 121–130.
Granger, C.W.J., Hallman, J.J., 1991. Nonlinear transformations of integrated time series. J. Time Ser. Anal. 12, 207–224.
Granger, C.W.J., Swanson, N., 1997. An introduction to stochastic unit root processes. J. Econ. 80, 35–62.
Hamilton, J.D., 1994. Time Series Analysis, Princeton University Press, Princeton.
Hansen, B.E., 2003. Testing for structural change in conditional means. J. Econ. 97, 93–115.
Hansen, B.E., Seo, B., 2002. Testing for two-regime threshold cointegration in vector correction models. J. Econ. 110, 293–318.
Jensen, J.L., Petersen, N.V., 1999. Asymptotic normality of the maximum likelihood estimator in state space models. Ann. Stat. 27, 514–535.
Johansen, S., 1988. Statistical analysis of cointegrating vectors. J. Econ. Dyn. Control 12, 231–254.
Johansen, S., 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica 59, 1551–1580.
Johansen, S., 1992. A representation of vector autoregressive processes integrated of order 2. Econ. Theory 8, 188–202.
Johansen, S., 1995. Likelihood-Based Inference in Cointegrated Vector Autoregressive Models, Oxford University Press, Oxford.
Juselius, K., 2006. The Cointegrated VAR Model: Methodologies and Applications, Oxford University Press, Oxford.
Kallianpur, G., Robbins, H., 1954. The sequence of sums of independent random variables. Duke Math. J. 21, 285–307.
Karlsen, H., Myklebust, T., Tjøstheim, D., 2007. Nonparametric estimation in a nonlinear cointegration type model. Ann. Stat. 35, 252–299.
Karlsen, H., Myklebust, T., Tjøstheim, D., 2010. Nonparametric regression estimation in a null recurrent time series. J. Stat. Plan. Inference 140, 3619–3626.
Karlsen, H., Tjøstheim, D., 2001. Nonparametric estimation in null recurrent time series models. Ann. Stat. 29, 372–416.
Kitagawa, G., 1987. Non-Gaussian state space modeling of nonstationary time series (with discussion). J. Am. Stat. Assoc. 82,1032–1063.
Koopman, S.J., Lucas, A., Klaassen, P., 2005. Empirical credit cycles and capital buffer formation. J. Bank. Financ. 29, 3159–3179.
Koopman, S.J., Ooms, M., 2006. Forecasting daily time series using periodic unobserved components time series models. Comput. Stat. Data Anal. 51, 885–903.
Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. J. Econ. 54, 159–178.
Leeb, H., Pötscher, B., 2001. The variance of an integrated process need not diverge to infinity, and related results on partial sums of stationary processes. Econ. Theory 17, 671–685.
Leybourne, S.J., McCabe, B., Mills, T.C., 1996. Randomized unit root processes for modeling and forecasting time series. J. Forecast. 15, 253–270.
Li, D., Ling, S., 2011. On the least squares estimation of multiple-regime threshold autoregressive models. Bernoulli 17.
Li, D., Ling, S., Tong, H., 2011. On moving-average models with feedbacks. Bernoulli 17 .

Liebscher, E., 2005. Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. J. Time Ser. Anal. 26, 669–691.
Ling, S., McAleer, M., 2002. Stationarity and the existence of moments of a family of GARCH models. J. Econ. 106, 109–117.
Ling, S., McAleer, M., 2010. A general asymptotic theory for time series models. Stat. Neerl. 64, 97–111.
Ling, S., Tong, H., 2005. Testing a linear MA model against threshold MA models. Ann. Stat. 33, 2529–2552.
Meitz, M., Saikkonen, P., 2008. Ergodicity, mixing and existence of moments of a class of Markov models with applications to GARCH and ACD models. Econ. Theory 24, 1291–1320.
Meitz, M., Saikkonen, P., 2011. Parameter estimation in nonlinear AR-GARCH models. Econ. Theory 27, 1236-1278.
Menkveld, A.J., Koopman, S.J., Lucas, A., 2007. Modelling round-the-clock price discovery for cross-listed stocks using state space methods. J. Bus. Econ. Stat. 25, 213–225.
Meyn, S.P., Tweedie, R.L., 1993. Markov Chains and Stochastic Stability, Springer, New York.
Myklebust, T., Karlsen, H.A., Tjøstheim, D., 2011. Null recurrent unit root processes. Econom. Theory 27.
Nicolau, J., 2002. Stationary processes that look like random walks. the bounded random walk process in discrete and continuous time. Econom. Theory 18, 99–118.
Ozaki, T., 1982. The statistical analysis of perturbed limit cycle processes using nonlinear time series models. J. Time Ser. Anal. 3, 29–41.
Ozaki, T., 1985. Non-linear time series models and dynamic systems. In: Hannan, E.J., Krishnaiah, P.R., Rao, M.M. (Eds.), Handbook in Statistics. Elsevier, Amsterdam.
Park, J.Y., Phillips, P.C.B., 1999. Asymptotics for nonlinear transformations of integrated time series. Econom. Theory 15, 269–298.
Park, J.Y., Phillips, P.C.B., 2001. Nonlinear regression with integrated time series. Econometrica 69, 117–161.
Phillips, A.W., 1957. Stabilization policy and the forms of lagged responses. Econ. J. 67, 265–277.
Phillips, P.C.B., 1987. Time series regression with a unit root. Econometrica 55, 277–301.
Phillips, P.C.B., 1991. Optimal inference in cointegrated systems. Econometrica 59, 283–306.
Phillips, P.C.B., Solo, V., 1992. Asymptotics for linear processes. Ann. Stat. 20, 971–1001.
Pötscher, B., 2002. Lower risk bounds and properties of confidence sets for ill-posed estimation problems with applications to spectral density and persistence estimation, unit roots and estimation of long memory parameters. Econometrica 70, 1035–1065.
Priestley, M.B., 1965. Evolutionary spectra and nonstationary processes. J. Time Ser. Anal. 27, 204–237.
Revuz, D., Yor, M., 1994. Continuous Martingale and Brownian Motion, second ed. Springer, New York.
Rico, V.B., Gonzalo, J., 2010. Summability of stochastic processes. a generalization of integration and co-integration valid for non-linear processes. Manuscript, Universidad Carlos III de Madrid.
Ripatti, A., Saikkonen, P., 2001. Vector autoregressive processes with nonlinear time trends in cointegrating relations. Macroecon. Dyn. 5, 577–597.
Saikkonen, P., 2005. Stability results for nonlinear error correction models. J. Econ. 129, 69–81.
Saikkonen, P., Choi, I., 2004. Cointegrating smooth transition regressions. Econom. Theory 20, 301–340.
Sancetta, A., 2009. Nearest neighbor conditional estimation for Harris recurrent Markov chains. J. Multivar. Anal. 100, 2224–2236.
Sargan, J.D., 1964. Wages and prices in the United Kingdom: a study in econometric methodology. In: Hart, P.E., Mills, G., Whittaker, J.N., (Eds.), Econometric Analysis for National Economic Planning, Butterworths, London.
Shephard, N.G. (Eds.), 2005. Stochastic Volatility. Selected Readings, Oxford University Press, Oxford.
Shephard, N.G., Pitt, M.K., 1997. Likelihood analysis of non-Gaussian measurement time series. Biometrika 84, 653–667.
Sperlich, S., Tjøstheim, D., Yang, L., 2002. Nonparametric estimation and testing of interaction in additive models. Econom. Theory 18, 197–251.
Straumann, D., 2005. Estimation in Conditionally Heteroscedastic Time Series Models, Lecture Notes in Statistics, Springer, New York.
Subba Rao, T., 1981. On the theory of bilinear models. J. R. Stat. Soc. Ser. B 43, 224–255.
Subba Rao, T., Gabr, M.M., 1984. An Introduction to Bispectral Analysis and Bilinear Time Series Models, Springer, New York.

Tanaka, K., 1996. Time Series Analysis Nonstationary and Noninvertible Distribution Theory, Wiley, New York.

Teräsvirta, T., Tjøstheim, D., Granger, C.W.J., 2010. Modelling Nonlinear Economic Time Series, Oxford University Press, Oxford.

Tjøstheim, D., 1986. Estimation in nonlinear time series models. Stoch. Process. Appl. 21, 251–273.

Tong, H., 1990. Non-linear Time Series. A Dynamical System Approach, Oxford University Press, Oxford.

Tong, H., Lim, K.S., 1980. Threshold autoregression, limit cycles and cyclical data (with discussion), J. R. Stat. Soc. Ser. B 42, 245–292.

Tsay, R.S., 1998. Testing and modeling threshold autoregressive processes. J. Am. Stat. Assoc. 93, 1188–202.

Wang, Q., Phillips, P.C.B., 2009a. Asymptotic theory for local time density estimation and nonparametric cointegrating regression. Econom. Theory 25, 710–738.

Wang, Q., Phillips, P.C.B., 2009b. Structural nonparametric cointegrating regression. Econometrica 77, 1901–1948.

Wang, Q., Phillips, P.C.B., 2010. Specification Testing for Nonlinear Cointegrating Regression, Manuscript, Department of Economics, Yale University.

Watson, M.W., 1994. Vector autoregression and cointegration. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4. North Holland, Amstardam, pp. 2844–2918.

Wong, C.S., Li, W.K., 2000. On a mixture of autoregressive models. J. R. Stat. Soc. Ser. B 62, 95–115.

Xia, Y., 1998. Doctoral Thesis, University of Hong Kong.

Yakowitz, S.J., 1993. Nearest neighbour regression estimation for null-recurrent Markov time series. J. Appl. Probab. 48, 311–318.

Yao, J.-F., Attali, J.-F., 2000. On stability of nonlinear AR processes with Markov switching. Adv. Appl. Probab. 32, 394–407.

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00005-3

Markov Switching Time Series Models

Jürgen Franke
University of Kaiserslautern, Department of Mathematics, Erwin-Schroedinger-Str., 67663 Kaiserslautern, Germany

Abstract

We give a review of time series with regime-switching, which look stationary over limited time intervals, but where the data-generating mechanism may suddenly change sometimes. After briefly discussing observation driven switching, we focus on Markov switching where changes are controlled by a hidden Markov chain. We illustrate the fundamental problems linked with such models, i.e., parameter estimation with only partly observable data and filtering, i.e., reconstruction of the hidden data, by looking at the simple, but nontrivial problem of Markov switching autoregressions in detail. In particular, we provide the details of the EM algorithm and the Viterbi algorithm as feasible solutions of the estimation and filtering problem. Additionally, we discuss references to more complex models like Markov switching ARMA and GARCH processes and to related continuous-time models from mathematical finance.

Keywords: regime switching, hidden Markov, filtering, autoregression, GARCH.

1. Introduction

This chapter is concerned with time series, which piecewise look like realizations from well-known simple stationary processes. Sometimes, however, the visual appearance of the data changes more or less abruptly to look again homogeneous, but differently, afterward. In the econometrics literature, this phenomenon is called regime switching (Franses and van Dijk, 2000; Lange and Rahbek, 2009). The main feature of regime-switching models is the existence of a typically finite number of states or regimes, represented by more or less simple data-generating mechanisms, between which the system changes repeatedly. Switches between the regimes are sudden or occur over

a very short period of time in contrast to data that are modeled by locally stationary processes where changes are more gradual. This kind of behavior is similar to that of time series with changepoints. In the latter case, however, usually it is assumed that the data-generating mechanism changes once for all at the changepoint, whereas the time series considered here will visit the various regimes over and over again.

Time series data exhibiting the described behavior can be found in many fields of application. Krolzig (1997) discusses such models in business cycle analysis, where the regimes correspond to various states of the economy. Guidolin and Timmermann (2007) apply regime-switching models to asset allocation where the regimes correspond to different states of the market like crash, slow growth, bull or recovery. Müller et al. (1995) and Liehr et al. (1999a) have a look at biological signals, more precisely at electroencephalograms recorded during sleep, which switch through different regimes representing various degrees of deep sleep, light slumber or dream phases. Tadjuidje et al. (2009) consider electroencephalograms in the presence of external stimuli to the brain. Peng et al. (1996) discuss modeling of speech signals, and Pinson et al. (2008) consider regime-switching models for wind time series.

There are various different ways to model a time series $\{X_t\}$ that switches through a finite number of regimes. The change of states is controlled by a switching variable Q_t which is a time series itself assuming only finitely many values, say $1, \ldots, K$, corresponding to the regime numbers. In the univariate case, a rather general model is given by

$$X_t = F(X_{t-1}, \ldots, X_{t-m}, Q_t, \epsilon_t; \theta), \tag{1}$$

where the innovations ϵ_t are independent identically distributed (i.i.d.) with known distribution and θ denotes the vector of model parameters. The distribution of X_t given the past of the process up to time $t-1$ is assumed to depend explicitly only on the last m observations which includes general nonlinear autoregressive and ARCH schemes, but not switching ARMA and GARCH models for which we give some references below.

Based on (1), there are two distinct approaches differing with respect to the dependence between the switching variable and the data of interest. One large class of models, which Lange and Rahbek (2009) call observation switching models, is based on the assumption that, given the data, the switching variable Q_t does not depend on past $Q_s, s < t$. More precisely, including the usual independence assumption for the innovation:

$$\begin{aligned}\mathbb{P}(Q_t = k, \epsilon_t \in B | Q_{t-1}, \ldots, Q_0, X_{t-1}, \ldots, X_0) \\ = \mathbb{P}(Q_t = k | X_{t-1}, \ldots, X_0)\ \mathbb{P}(\epsilon_t \in B).\end{aligned}$$

Observation switching models include the popular threshold models of Tong (1990). The simple example of a self-exciting threshold autoregression (SETAR) of order 1 with only two regimes

$$X_t = \begin{cases} \alpha_1 X_{t-1} + \sigma\epsilon_t & \text{if } X_{t-1} \le c, \\ \alpha_2 X_{t-1} + \sigma\epsilon_t & \text{else,} \end{cases}$$

is obviously of the form (1) with $\theta = (\alpha_1, \alpha_2, \sigma, c)^T$ and $Q_t = 1$ if $X_{t-1} \le c$ and $= 2$ otherwise.

More complicated observation switching models and in particular their application to financial data are extensively discussed in the monograph of the works done by Franses and van Dijk (2000). In this survey, however, we focus on the second large class of time series with regime switching, the Markov switching models, which generalize the hidden Markov models. Here, the switching variables Q_t form a Markov chain with finite state space, and the conditional distribution of Q_t given the past up to time $t-1$ depends only on Q_{t-1} and not on the data $X_{t-1}, \ldots, X_0$. We give a precise formulation of that crucial assumption below.

The terminology regarding models with Markovian switching variables is not completely standardized in the literature. We follow Cappé et al. (2005) and call a regime-switching time series a hidden Markov process if the temporal dependence of the observations is completely induced by the Markovian structure of the switching variables, i.e., the conditional distribution of X_t given $Q_t, Q_s, X_s, s < t$, depends on Q_t only and, in particular, X_t is independent of $X_s, s < t$, conditional on Q_t. For the special case of (1), a hidden Markov model is given by $X_t = F(Q_t, \epsilon_t; \theta)$.

Regime-switching models for time series have not only found a lot of interest in statistics during the last two decades, but also in machine learning. There, such models are frequently called mixtures-of-experts (compare, e.g., Jiang and Tanner (1999) and Liehr et al. (1999a)). In particular, Carvalho and Tanner (2005) give a detailed analysis including asymptotics and a discussion of model selection for mixtures of autoregressive processes, where the switching between regimes is driven by the observation. However, the terminology is not always consistent as some authors like Liehr et al. (1999b) also call models with hidden switching variables mixtures-of-experts.

The literature on hidden Markov models and Markov switching time series models is now quite extensive. Here, we only give an introduction to the main ideas by having a detailed look at a simple, but nontrivial example in the following. We consider Markov switching autoregressions of order 1 or MS-AR(1) with only two different regimes. This example already exhibits the main features of Markov switching models, but allows for a still easily accessible notation. We discuss the theoretical properties and numerical calculation of estimates for the model parameters as well as solutions to the filtering problem, i.e., the reconstruction of the hidden sequence Q_t from the observations. For more details and applications, we refer to the excellent monographs of MacDonald and Zucchini (1997), Cappé et al. (2005), and Frühwirth-Schnatter (2006). We close with a short review of other popular Markov switching time series models like Markov switching ARCH and GARCH and have a look at a corresponding class of models for continuous-time data, mainly from finance.

2. Markov switching autoregressions

We start from a univariate time series $\{X_t\}$ that is influenced by a hidden Markov chain $\{Q_t\}$ assuming only finitely many values $1, \ldots, K$. The current value Q_t represents the state or regime of the mechanism generating the data X_t. For a concise notation, we write $S_{tk} = 1$ if $Q_t = k$, and $= 0$ else. The sequence of unit vectors $S_t = (S_{t1}, \ldots, S_{tK})$ provides an equivalent representation of the state variables Q_t.

We assume that the single regimes are all characterized by linear autoregressions with varying orders m_k, parameters $\alpha_{k,1}, \ldots, \alpha_{k,m_k}$, and innovation variances σ_k^2. The

whole process $\{X_t\}$ is, then, given by the following mixture of autoregressions and usually called a Markov switching autoregression (MS-AR):

$$X_t = \sum_{k=1}^{K} S_{tk} \left(\sum_{s=1}^{m_k} \alpha_{k,s} X_{t-s} + \sigma_k \, \epsilon_t \right). \tag{2}$$

The innovations ϵ_t are assumed to be i.i.d. with mean 0 and variance 1. Here, we only consider the case where ϵ_t has a density $p_\epsilon(u) > 0$ for all $u \in \mathbb{R}$.

The distribution of the hidden state process is given by the $K \times K$ transition probability matrix A, i.e.,

$$A_{jk} = \mathbb{P}(Q_t = k | Q_{t-1} = j),$$

and we denote the weights of the corresponding stationary distribution by $\pi = (\pi_1, \ldots, \pi_K)$, i.e., in the stationary state we have $\pi_k = \mathbb{P}(Q_t = k)$. Mark that the latter are determined by A via $\pi A = \pi$.

Markov switching autoregressions have been proposed by Hamilton (1989, 1990) in econometrics and since then become quite popular. They build upon earlier work on hidden Markov models like that of Baum and Petrie (1966) or Lindgren (1978), and have been extended by Holst et al. (1994) and McCulloch and Tsay (1994). As the Markovian as well as the autoregressive structure introduce temporal dependence into the model, the autocorrelation structure is quite flexible. Timmermann (2000) has given explicit formulas for the autocorrelations of stationary Markov switching autoregressions.

To keep the notation as clear as possible, we simplify the model further by considering mainly the case of only two regimes ($K = 2$), both represented by first-order autoregressions ($m_1 = m_2 = 1$):

$$X_t = \begin{cases} \alpha_1 X_{t-1} + \sigma_1 \epsilon_t & \text{if } Q_t = 1, \\ \alpha_2 X_{t-1} + \sigma_2 \epsilon_t & \text{if } Q_t = 2. \end{cases} \tag{3}$$

For given p_ϵ, model (3) is characterized by the parameter vector

$$\theta = \left(\alpha_1, \alpha_2, \sigma_1^2, \sigma_2^2, A_{11}, A_{22}\right)^T.$$

The stationary regime probabilities are given by $\pi_1 = A_{12}/(A_{12} + A_{21})$ and $\pi_2 = 1 - \pi_1$. Frequently, we consider the specific case where the innovations ϵ_t are standard normal variables. In the following, $\varphi(\cdot; \mu, \sigma^2)$ denotes the density of the normal law with mean μ and variance σ^2. Then, the conditional distribution of X_t given X_{t-1} and $Q_t = k$ has the density $\varphi(\cdot; \alpha_k X_{t-1}, \sigma_k^2)$. Other distributions with positive densities may be considered analogously.

Figure 1 shows a realization of a time series satisfying (3) with parameters $\alpha_1 = 0.9, \alpha_2 = -0.9, \sigma_1^2 = 1, \sigma_2^2 = 0.25, A_{11} = 0.8, A_{22} = 0.9,$ and $N = 200$. Figure 2 shows the scatter plot $(X_{t-1}, X_t), t = 1, \ldots, N$. Such scatter plots are, in general, a good exploratory tool for detecting regime switching, as they frequently differ in one

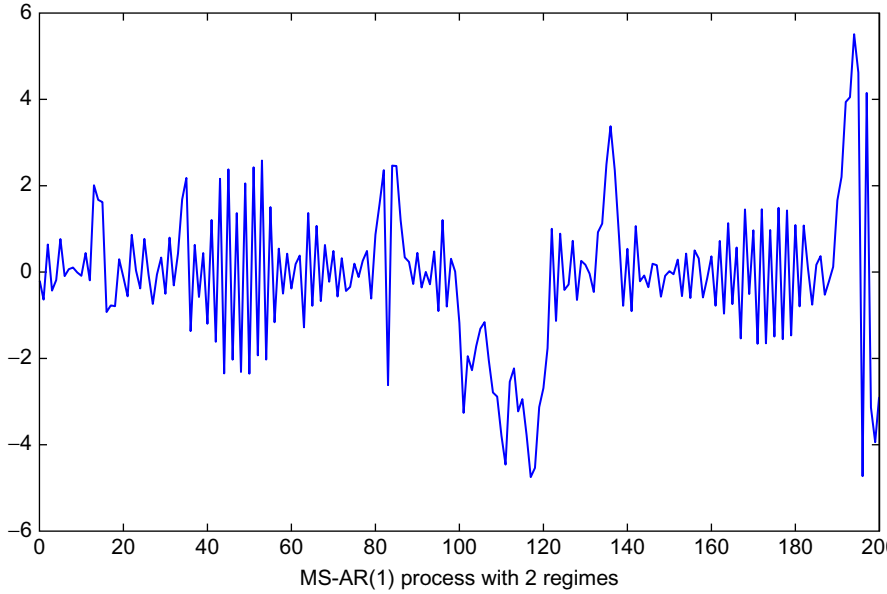

Fig. 1. Markov switching autoregressions of order 1 with two regimes.

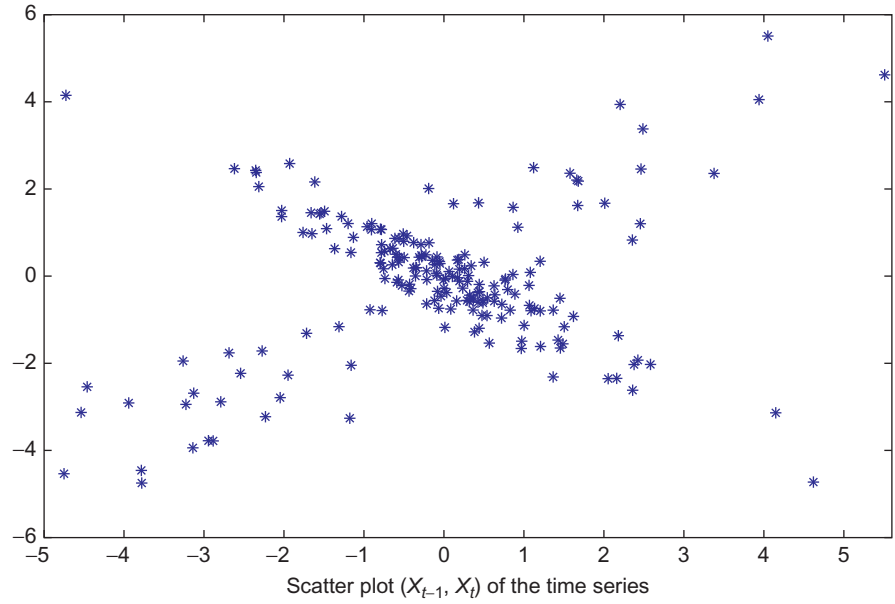

Fig. 2. Scatter plot of the data shown in Fig. 1.

way or the other from the elliptic shape known from Gaussian or similar linear time series. Of course, the parameter constellation of this example is rather extreme such that the effect is rather pronounced.

Model (3) also covers the case of an autoregression of order 1 with a single changepoint in the classical sense. If we choose

$$A = \begin{pmatrix} A_{11} & 1 - A_{11} \\ 0 & 1 \end{pmatrix}$$

then the time series will never return to regime 1 once it has changed to regime 2 (compare Bauwens and Rombouts (2010) for the case of multiple changepoints). In spite of the formal similarity, there is an important difference between the setup of changepoint analysis and a genuine Markov switching time series with finitely many regimes, which is recurrent in the sense that it returns to every regime over and over again. Later on, this feature will be guaranteed by assuming that A_{11}, A_{22} are bounded away from 0 and 1. Then, for sample size $N \to \infty$, there will be an increasing number of observations from every regime that allows for estimating all the parameters consistently. To get this essential feature in the situation of a single changepoint, we have to let $A_{11} \to 1$ with $N \to \infty$ to have an increasing number of data before and after the changepoint, i.e., from both regimes.

In the next subsection, we discuss estimating the parameters of model (3). If we could observe the state process Q_t, this would be an easy exercise. The main difficulty which is characteristic for Markov switching time series models arises from the fact that this Markov chain is hidden. To illustrate those problems, we first have a look at an even simpler special case that corresponds to the mixture autoregressive model of Wong and Li (2000). Here, the Q_t are i.i.d. with $\pi_1 = \mathbb{P}(Q_t = 1), \mathbb{P}(Q_t = 2) = \pi_2 = 1 - \pi_1$. The transition matrix is given by

$$A = \begin{pmatrix} \pi_1 & \pi_2 \\ \pi_1 & \pi_2 \end{pmatrix}.$$

The model parameters are now $\theta = (\alpha_1, \alpha_2, \sigma_1^2, \sigma_2^2, \pi_1)^T$. We consider observations $X_0, \ldots, X_N$, whereas the state variables Q_t are unknown. The conditional probability density of a single observation X_t at y given $X_{t-1} = x$ is

$$p_\theta(y|x) = \sum_{k=1}^{2} \pi_k \varphi\big(y; \alpha_k x, \sigma_k^2\big).$$

As usual for stationary time series with an autoregressive dependence (Brockwell and Davis, 1991), we consider the conditional log-likelihood given the initial observation X_0. In the case of i.i.d. state variables Q_t, this is given by

$$\ell(\theta|\mathbf{X}^{(N)}) = \sum_{t=1}^{N} \log p_\theta(X_t|X_{t-1}),$$

where $\mathbf{X}^{(N)} = (X_0, \ldots, X_N)^T$. If, on the other hand, the state variables $\mathbf{Q}^{(N)} = (Q_0, \ldots, Q_N)^T$ would be observable, the so-called complete likelihood would be of

a much simpler form, using that the state indicators S_{tk} assume the values 0 and 1 only:

$$\ell_c(\theta|\mathbf{X}^{(N)}, \mathbf{Q}^{(N)}) = \sum_{t=1}^{N}\sum_{k=1}^{2} S_{tk} \log\left[\pi_k \varphi\left(X_t; \alpha_k X_{t-1}, \sigma_k^2\right)\right]$$

$$= \sum_{t=1}^{N}\sum_{k=1}^{2} S_{tk} \log \pi_k + \ell_c^{AR}(\theta|\mathbf{X}^{(N)}, \mathbf{Q}^{(N)}),$$

which consists of two parts, the first depending only on the state variable parameters, and the second depending only on the autoregressive parameters. The latter is given by

$$\ell_c^{AR}(\theta|\mathbf{X}^{(N)}, \mathbf{Q}^{(N)}) = -\frac{1}{2}\sum_{t=1}^{N}\sum_{k=1}^{2} S_{tk}\left(\log \sigma_k^2 + \frac{(X_t - \alpha_k X_{t-1})^2}{\sigma_k^2}\right) + \text{const.} \tag{4}$$

ℓ_c may be maximized explicitly by setting the derivatives with respect to the parameters to 0, and we get straightforwardly:

$$\hat{\pi}_1 = \frac{1}{N}\sum_{t=1}^{N} S_{t1} = 1 - \hat{\pi}_2, \quad \hat{\alpha}_k = \frac{\sum_{t=1}^{N} S_{tk} X_t X_{t-1}}{\sum_{t=0}^{N-1} S_{tk} X_t^2},$$

$$\hat{\sigma}_k^2 = \frac{1}{N\hat{\pi}_k}\sum_{t=1}^{N} S_{tk}(X_t - \hat{\alpha}_k X_{t-1})^2. \tag{5}$$

ℓ, however, has to be maximized numerically as p_θ is not of exponential form. A popular algorithm with a statistical intuition is the EM algorithm (compare, e.g., Dempster et al. (1977) and Wu (1983)). In the M-step, it replaces the hidden variables S_{tk} in (5) by their conditional expectations given $X_0, \ldots, X_N$, and, in the E-Step, it calculates those conditional expectations using the current estimates of the model parameters. Then, both steps are iterated. For general nonlinear autoregressions, Franke et al. (2011) have shown convergence of this algorithm and consistency of the resulting estimates using kernel smoothers for estimating the autoregressive function. The details will be discussed below in the more general context of Markov switching.

2.1. *Maximum likelihood estimates*

We now derive the log-likelihood function $\ell(\theta|\mathbf{X}^{(N)})$ of model (3) for the general case where, again, we always condition on the initial observation X_0.

We denote by $\mathbf{x}^{(N)} = (x_0, \ldots, x_N)^T, \mathbf{q}^{(N)} = (q_0, \ldots, q_N)^T$ possible values of $\mathbf{X}^{(N)}$, $\mathbf{Q}^{(N)}$. We characterize the joint distribution of a part Y of $\mathbf{X}^{(N)}$ and a part Z of $\mathbf{Q}^{(N)}$ by its density $p_\theta(y, z)$ with respect to the product measure $\lambda \otimes \nu$, where λ denotes the Lebesgue measure and ν denotes the counting measure on the set of possible values of Z, i.e.,

$$\mathbb{P}(Y \in B, Z \in C) = \sum_{z \in C} \int_B p_\theta(y, z) dy.$$

Looking at a single observation $Y = X_t$, a crucial feature of Markov switching first-order autoregressions, obvious from (3), is given by

$$p_\theta\left(y \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}, \mathbf{Q}^{(t)} = \mathbf{q}^{(t)}\right) = p_\theta(y \mid X_{t-1} = x_{t-1}, Q_t = q_t), \tag{6}$$

i.e., the current observation depends only on the current state Q_t and on the previous observation. In more complex Markov switching models, X_{t-1} is replaced by several random variables from the past of the process up to time $t-1$.

Another crucial feature of Markov switching models is the assumption that the evolution of the Markov chain in time does not depend on the observations:

(A1) Let $\mathcal{F}_s = \sigma(X_t, t \le s)$ denote the σ-algebra generated by the observations up to time s, and let B_{t-1} be any event in $\mathcal{F}_{t-1}$. Then,

$$\mathbb{P}(Q_t = j | Q_{t-1} = i, B_{t-1}) = \mathbb{P}(Q_t = j | Q_{t-1} = i) \quad \text{for all } i, j.$$

The joint density of $\mathbf{X}^{(N)}, \mathbf{Q}^{(N)}$ is given by

$$p_\theta\left(\mathbf{x}^{(N)}, \mathbf{q}^{(N)}\right) = p_\theta\left(\mathbf{x}^{(N)} \mid \mathbf{Q}^{(N)} = \mathbf{q}^{(N)}\right) p_A\left(\mathbf{q}^{(N)}\right),$$

where $p_A\left(\mathbf{q}^{(N)}\right) = \mathbb{P}\left(\mathbf{Q}^{(N)} = \mathbf{q}^{(N)}\right)$ does not depend on the autoregressive part of the parameter vector θ, but only on the parameters A_{11}, A_{22} of the transition matrix A. As Q_t is Markovian and if we start it with its stationary distribution at time 0, we have for the latter

$$p_A\left(\mathbf{q}^{(N)}\right) = \left(\prod_{t=1}^{N} p_A\left(q_t | Q_{t-1} = q_{t-1}\right)\right) \mathbb{P}(Q_0 = q_0) = \pi_{q_0} \prod_{t=1}^{N} A_{q_{t-1}, q_t}.$$

From (A1) and (6), we get

$$\begin{aligned} p_\theta\left(\mathbf{x}^{(N)}, q^{(N)}\right) &= p_\theta\left(x_N \mid \mathbf{X}^{(N-1)} = \mathbf{x}^{(N-1)}, \mathbf{Q}^{(N)} = q^{(N)}\right) \\ &\quad \times p_\theta\left(q_N | \mathbf{X}^{(N-1)} = \mathbf{x}^{(N-1)}, \mathbf{Q}^{(N-1)} = q^{(N-1)}\right) \\ &\quad \times p_\theta\left(\mathbf{x}^{(N-1)}, q^{(N-1)}\right) \\ &= p_\theta\left(x_N \mid X_{N-1} = x, Q_N = q_N\right) p_\theta(q_N | Q_{N-1} = q_{N-1}) \\ &\quad \times p_\theta\left(\mathbf{x}^{(N-1)}, q^{(N-1)}\right) \end{aligned} \tag{7}$$

Iterating (7) we immediately get for the complete likelihood conditional on X_0

$$L_c\left(\theta | \mathbf{X}^{(N)}, \mathbf{Q}^{(N)}\right) = \pi_{Q_0} \prod_{t=1}^{N} A_{Q_{t-1}, Q_t} p_\theta(X_t \mid X_{t-1}, Q_t)$$

and for the corresponding log-likelihood, assuming Gaussian innovations,

$$\ell_c\left(\theta | \mathbf{X}^{(N)}, \mathbf{Q}^{(N)}\right) = \log \pi_{Q_0} + \sum_{t=1}^{N} \log A_{Q_{t-1}, Q_t} + \ell_c^{AR}\left(\theta | \mathbf{X}^{(N)}, \mathbf{Q}^{(N)}\right),$$

where ℓ_c^{AR} is the same as in (4). As the state variables Q_t are hidden, we have to consider the incomplete likelihood that we get by summing over all possible paths of the Markov chain starting at time $t = 0$:

$$\begin{aligned} L\left(\theta|\mathbf{X}^{(N)}\right) &= \sum_{q_0,\ldots,q_N=1}^{2} \pi_{q_0} \prod_{t=1}^{N} A_{q_{t-1},q_t}\, p_\theta(X_t \mid X_{t-1}, Q_t = q_t) \\ &= \sum_{q_0,\ldots,q_N=1}^{2} \pi_{q_0} \prod_{t=1}^{N} A_{q_{t-1},q_t} \left(\sum_{k=1}^{2} s_{tk}\varphi\left(z;\alpha_k x, \sigma_k^2\right)\right) \end{aligned} \tag{8}$$

with $s_{tk} = 1$ if $q_t = k$, and $= 0$ else. To get the maximum likelihood estimate $\hat{\theta}_N$ of the model parameter θ, we have to maximize $\ell\left(\theta|\mathbf{X}^{(N)}\right) = \log L\left(\theta|\mathbf{X}^{(N)}\right)$ numerically. Before we consider suitable algorithms, we have a look at the asymptotic properties of those estimates in the next sections.

2.2. *Ergodicity and consistency*

The key to inference about Markov switching time series models is the ergodicity of the corresponding stochastic process. In addition to (A1) above, the following assumptions about the hidden Markov chain and its dependence on the observed data are needed:

- **(A2)** The process $\{Q_t\}$ is a strictly stationary, irreducible, and aperiodic Markov chain.
- **(A3)** The innovations ε_t are i.i.d. with mean 0 and variance 1 and have a continuous, positive density.
- **(A4)** ε_t is independent of $Q_t, X_{t-1}, X_{t-2}, \ldots$

Such assumptions are more or less generic. They may be relaxed to some extent but appear in a similar form without regard to the particular time series models that represent the various regimes. Additionally, some assumptions about the structure of the individual mixture components are needed. In the simple model (3), those assumptions reduce to

- **(A5)** $\sum_{k=1}^{2} A_{lk}\alpha_k^2 < 1 \quad \text{for} \quad l = 1, 2.$

THEOREM 1. *Let $\{X_t\}$ follow (3), and let Q_t, X_t, ε_t satisfy assumptions (A1)–(A5). Then, the joint process $(S_t, X_t)^T$ is geometrically ergodic.*

This result is a special case of Theorem 2.1 of Franke et al. (2010), where mixtures of general conditionally heteroscedastic autoregressions are considered. A similar result for nonlinear Markov switching autoregressions has been shown by Yao and Attali (2000) who also prove the existence of a stationary solution and a strong law of large numbers.

If both regimes in (3) correspond to stationary AR(1)-processes, i.e., if $|\alpha_k| < 1$, $k = 1, 2$, then, due to $A_{l1} + A_{l2} = 1$, condition (A5) will be automatically satisfied. An interesting feature of Markov switching models is, however, that not all component models have to satisfy stationarity conditions. Some of them may correspond to explosive processes, provided that they show up rarely enough in the

course of the combined process. For example, consider (3) with $|\alpha_2| > 1 > |\alpha_1|$. Then, condition (A5) is equivalent to requiring

$$A_{11} > \frac{\alpha_2^2 - 1}{\alpha_2^2 - \alpha_1^2}, \quad A_{22} < \frac{1 - \alpha_1^2}{\alpha_2^2 - \alpha_1^2}.$$

For example, $\alpha_1^2 = 0.5, \alpha_2^2 = 1.5$, (A5) holds for $A_{11} > 0.5 > A_{22}$ and a straightforward calculation shows $\pi_1 > 0.5$. Therefore, a stationary solution of (3) may exist if the first AR(1) component satisfies the usual stationarity condition $|\alpha_1| < 1$ and the second one corresponds to an explosive regime. A necessary consequence is that the whole process X_t has to spend more time in the stationary regime than in the explosive one.

Francq and Roussignol (1998) consider general nonlinear autoregressive mixtures of order 1 which, in line with our notation, may be written as

$$X_t = \sum_{k=1}^{K} S_{tk}\{F_k(X_{t-1}; \theta) + G_k(\varepsilon_t; \theta)\}.$$

The data X_t and the innovations ε_t may be multivariate such that this model also covers higher order univariate autoregressions via their vector state-space representations. They prove under appropriate assumptions the existence of an ergodic stationary solution and, then, the strong consistency of the maximum likelihood estimator $\hat{\theta}_N$ of θ. Mark that the transition probabilities $A_{ij} = A_{ij}(\theta)$ of the underlying Markov chain depend on the general parameter θ. In the simple case of (3), θ consists of the autoregressive parameters $\theta = (\alpha_1, \alpha_2, \sigma_1^2, \sigma_2^2, A_{11}, A_{22})^T$. For (3), Francq and Roussignol (1998) have given a detailed analysis of the case where $\sigma_1^2 = \sigma_2^2 = \sigma^2$ does not depend on the regime. To remain within their framework, we consider only the case of standard normal innovations. Additionally, we assume

(A6) $\alpha_1 \neq \alpha_2, 0 < A_{22} < A_{11} < 1,$

where the latter also implies $\pi_1 > \pi_2$, i.e., regime 1 is the more frequently visited one. Identifiability assumptions like (A6) are an inherent feature of inference for Markov switching models. Without them, the model parameters are not uniquely determined by the distribution. The possibility of permuting the regimes, i.e., of changing their enumeration, without changing the data-generating mechanism, is always present. In our simple example, the parameters $(\alpha_1, \alpha_2, \sigma_1^2, \sigma_2^2, A_{11}, A_{22})$ and $(\alpha_2, \alpha_1, \sigma_2^2, \sigma_1^2, A_{22}, A_{11})$ give rise to the same stochastic process. In more complicated models, more involved identifiability conditions are needed, e.g., for the general case in the works done by Francq and Roussignol (1998) or for nonlinear autoregressive-ARCH models based on neural networks (Stockis et al., 2008).

Applying the same kind of argument as in Section 5 of the works done by Francq and Roussignol (1998), using that for any compact subset of the parameter set $\underline{\sigma}^2 \leq \sigma_1^2, \sigma_2^2 \leq \overline{\sigma}^2$ for some $\overline{\sigma}^2 \geq \underline{\sigma}^2 > 0$, we get

THEOREM 2. *Let $\{X_t\}$ follow (3), let ε_t be standard normal, and let Q_t, X_t, ε_t satisfy assumptions (A1)–(A4). Let θ_0 denote the true model parameter, and assume that*

it satisfies (A5) and (A6). Then, for any compact subset Θ^* of the parameter set containing θ_0 in its interior,

$$\hat{\theta}_N = \arg\max_{\theta\in\Theta^*} L(\theta|\mathbf{X}^{(N)}) \longrightarrow \theta_0 \quad a.s.$$

2.3. Asymptotic normality

A detailed discussion of the asymptotic properties of maximum likelihood estimates like $\hat{\theta}_N$ has been given by Douc et al. (2004). They consider models of the form

$$X_t = F(X_{t-1}, \ldots, X_{t-m}, Q_t, \epsilon_t; \theta), \tag{9}$$

(compare also (1) above), where the hidden Markov chain Q_t does not necessarily assume only finitely many values, but may have an arbitrary compact range. Douc et al. prove consistency of the maximum likelihood estimate $\hat{\theta}_N$ of θ conditional on Q_0 and $X_0, \ldots, X_{1-m}$, they show asymptotic negligibility of those initial values, and they prove asymptotic normality of $\hat{\theta}_N$ under the usual kind of regularity assumptions which are easily checked for the simple model (3) with Gaussian innovations. In addition to our previous assumptions, we have to restrict our attention from the outset to a compact parameter set Θ^* characterized by

(A7) $|\alpha_1|, |\alpha_2| \le \alpha^*, \underline{\sigma}^2 \le \sigma_1^2, \sigma_2^2 \le \overline{\sigma}^2, \delta \le A_{11}, A_{22} \le 1-\delta,$

for some suitable constants $\alpha^* < \infty, 0 < \underline{\sigma}^2 \le \overline{\sigma}^2 < \infty, 0 < \delta < 0.5$. With $I(\theta)$ denoting the Fisher information matrix, we get from Theorem 4 of Douc et al.

THEOREM 3. *Let the assumptions of Theorem 2 and, additionally, (A7) be satisfied. If the true model parameter θ_0 lies in the interior of Θ^* and if $I(\theta_0)$ is positive definite, then*

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \underset{d}{\longrightarrow} \mathcal{N}(0, I^{-1}(\theta_0)).$$

To get a representation of $I(\theta_0)$, following Douc et al., we first consider the conditional score of (Q_t, X_t) given (Q_{t-1}, X_{t-1})

$$\psi(\theta, Q_t, X_t, Q_{t-1}, X_{t-1}) = \nabla_\theta \log\left[A_{Q_{t-1},Q_t}\, p_\theta(X_t|X_{t-1}, Q_{t-1})\right],$$

where ∇_θ denotes the gradient with respect to $\theta = (\alpha_1, \alpha_2, \sigma_1^2, \sigma_2^2, A_{11}, A_{22})^T$ and

$$p_\theta(y|x,k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(y-\alpha_k x)^2}{2\sigma_k^2}\right]$$

is the conditional density of X_t given $X_{t-1} = x, Q_t = k$. Recalling $A_{12} = 1 - A_{11}$, $A_{21} = 1 - A_{22}$, it is a straightforward exercise to calculate ψ. With $\mathcal{F}_{-\infty}^s$ denoting the

σ-algebra generated by the data $X_s, X_{s-1}, \ldots$, we set

$$\Delta_t(\theta_0) = \mathbb{E}_{\theta_0}\left\{\psi(\theta_0, Q_t, X_t, Q_{t-1}, X_{t-1})|\mathcal{F}_{-\infty}^{t}\right\} + \sum_{s=-\infty}^{t-1}\Big[\mathbb{E}_{\theta_0}\left\{\psi(\theta_0, Q_s, X_s, Q_{s-1}, X_{s-1})|\mathcal{F}_{-\infty}^{t}\right\} - \mathbb{E}_{\theta_0}\left\{\psi(\theta_0, Q_s, X_s, Q_{s-1}, X_{s-1})|\mathcal{F}_{-\infty}^{t-1}\right\}\Big],$$

and finally we have

$$I(\theta_0) = \mathbb{E}_{\theta_0}\left[\Delta_0(\theta_0)\Delta_0^T(\theta_0)\right].$$

Theorem 3 of Douc et al. together with the consistency of the maximum likelihood estimate $\hat{\theta}_N$ also provides a consistent estimate of $I(\theta_0)$.

2.4. *Model selection*

A crucial problem in the application of Markov switching models to time series data is the choice of the number K of regimes. Rydén (1995) applies classical order selection criteria like AIC and BIC to the selection of regimes in hidden Markov models. In the same spirit, MacKay (2002) considers alternative penalized minimum distance procedures for selecting K. A detailed survey including a theoretical analysis of penalized maximum likelihood order selection is given in Chapter 15 of Cappé et al. (2005). They also discuss the relation to general likelihood ratio testing, for which regime-switching models are a particular challenge as the parameters of a regime become immediately nonidentifiable under the hypothesis that this regime is not necessary for modeling the data. That issue is discussed in some detail in Section 4 of Lange and Rahbek (2009). Chopin (2007) considers modified hidden Markov models, where the regimes are numbered in order of their appearance in the time series sample. This allows for a sequential Monte Carlo algorithm solving the problem of selecting the number of regimes and estimating the parameters simultaneously.

In general Markov switching models, the model selection problem becomes even more complicated as, in addition to the number of regimes, the complexity of the single regime models has to be determined, e.g., the autoregressive orders $m_1, \ldots, m_K$ in model (2). For such Markov switching autoregressions, Psaradakis and Spagnolo (2003, 2006) present Monte Carlo studies for illustrating successive and simultaneous choices of the number of regimes and the autoregressive orders. Zhang and Stine (2001) show how to use the sample autocovariances of hidden Markov models and Markov switching autoregressions to get lower bounds on the number of regimes. Frühwirth-Schnatter (2004) discusses the importance of selecting the number of regimes and the orders of autoregressions simultaneously, as, otherwise, there may be a tendency to choose too few regimes and too large orders.

2.5. The EM algorithm

As we have seen in Section 2.1, even in simple Markov switching models like (3) we have to determine the maximum likelihood estimates of the parameters numerically. A popular algorithm with an appealing statistical intuition is the expectation–maximization (EM) algorithm of Dempster et al. (1977), which has been further investigated by Wu (1983). It has already been applied to hidden Markov models by Baum et al. (1970) without using that name. For Markov switching autoregressions, EM algorithms have been proposed and investigated by Hamilton (1990) and Holst et al. (1994). We restrict the following discussion to the application of the EM algorithm to calculate the maximum likelihood estimates of Markov switching parameters. There are, however, other numerical procedures for handling this task. Of course, one could apply any decent optimization routine to maximizing the likelihood (8), and, for small sample sizes, that works reasonably well. However, special statistically motivated approaches show a more stable performance. Apart from the class of EM-like algorithms, Markov chain Monte Carlo algorithms are popular tools for parameter estimation in Markov switching models, compare, e.g., the study by Frühwirth-Schnatter (2001) or the monograph of Frühwirth-Schnatter (2006). Rydén (2008) presents a comparative study of both types of algorithms for a selection of regime-switching models.

The principle of the EM algorithm for Markov switching time series is an alternation between estimating the hidden variables Q_t given the model parameters, and estimating the model parameters given the hidden variables. The second or M-step exploits the simpler form of the complete likelihood $L_c(\theta \mid \mathbf{X}^{(N)}, \mathbf{Q}^{(N)})$ compared to the more complex incomplete likelihood $L(\theta \mid \mathbf{X}^{(N)})$. For the first or E-step, we have to solve the filtering problem that consists in reconstructing the hidden variables Q_t from the data.

2.5.1. Forward–backward procedure or E-step

In this subsection, we assume that the data $X_0, \ldots, X_N$ are generated by (3) with a stationary Markov chain, i.e., in particular, we have $\mathbb{P}(Q_t = i) = \pi_i, i = 1, 2, t \geq 0$. Moreover, we assume that the autoregressive parameters $\alpha_1, \alpha_2, \sigma_1^2, \sigma_2^2$ as well as the transition matrix A are known. i.e., $\theta = (\alpha_1, \alpha_2, \sigma_1^2, \sigma_2^2, A_{11}, A_{22})^T$ is given. As above, p_θ denotes the density of the observed variables with respect to Lebesgue measure or the joint density of observed and hidden variables with respect to the Lebesgue measure and the counting measure. To stress which of the hidden variables is considered, we write, e.g., $p(\mathbf{x}^{(N)}, Q_t = i)$ for the joint density of the whole observed sample $\mathbf{X}^{(N)}$ and the single hidden variable Q_t evaluated at $(\mathbf{x}^{(N)}, i) \in \mathbb{R}^{N+1} \times \{1, 2\}$.

2.5.1.1. Forward procedure

Let

$$v_i^t = p_\theta(X_0, \ldots, X_t, Q_t = i) = p_\theta(X_0, \ldots, X_t | Q_t = i)\pi_i,$$

where $p_\theta(\mathbf{x}^{(t)} | Q_t = i)$ is the conditional density of $\mathbf{X}^{(t)}$ given $Q_t = i$. Based on the assumptions (A1) and (A2) above, we get the following recursive scheme for

calculating v_i^t:

$$\begin{aligned} v_j^{t+1} &= p_\theta(X_0, \ldots, X_t, X_{t+1}, Q_{t+1} = j) \\ &= p_\theta(X_{t+1}|\mathbf{X}^{(t)}, Q_{t+1} = j) \sum_{i=1}^{2} \mathbb{P}(Q_{t+1} = j|\mathbf{X}^{(t)}, Q_t = i)\ p_\theta(\mathbf{X}^{(t)}, Q_t = i) \\ &= p_\theta(X_{t+1}|X_t, Q_{t+1} = j) \sum_{i=1}^{2} \mathbb{P}(Q_{t+1} = j|Q_t = i)\ p_\theta(\mathbf{X}^{(t)}, Q_t = i\ ) \\ &= b_j^{t+1} \left[\sum_{i=1}^{2} A_{ij} v_i^t \right] \end{aligned} \tag{10}$$

for $t = 1, \ldots, N-1$, where, using that the innovations ϵ_t are standard normal variables,

$$b_j^t = p(X_t|X_{t-1}, Q_t = j) = \varphi(X_t; \alpha_j X_{t-1}, \sigma_j^2). \tag{11}$$

As we always condition in X_0 and assume it to be given, and as Q_1 follows the stationary distribution π of the Markov chain, we get the initial condition for the recursion

$$v_j^1 = p_\theta(X_0, X_1, Q_1 = j) = \pi_j b_j^1.$$

This recursive calculation is called the *forward procedure*. Mark that, by summation over the states at the end $t = N$, we get the density of the whole sequence of observations $p_\theta(\mathbf{x}^{(N)}) = v_1^N + v_2^N$ which, using the forward procedure, can be calculated in a number of steps increasing linearly with N. This is a pleasant surprise as a look at, e.g., the likelihood (8) with its summation over all possible paths of $Q_0, \ldots, Q_N$ would have rather led us to expect a factor 2^N in the number of computations, i.e., an exponential increase in computational complexity.

2.5.1.2. Backward procedure

Analogously, we define the backward variable w_i^t as the conditional density of observing the future $X_s, s = t+1, \ldots, N$, given the present state i and the observation X_t at time t

$$w_i^t = p_\theta(X_{t+1}, \ldots, X_N|X_t, Q_t = i).$$

Using again (A1) and (A2), we get the recursion called the *backward procedure*

$$w_i^t = \sum_{j=1}^{2} p_\theta(X_{t+1}, \ldots, X_N, Q_{t+1} = j|X_t, Q_t = i)$$

$$
\begin{aligned}
&= \sum_{j=1}^{2} p_\theta(X_{t+2}, \ldots, X_N | X_{t+1}, Q_{t+1} = j) \\
&\quad \times p_\theta(X_{t+1} | X_t, Q_{t+1} = j) \mathbb{P}(Q_{t+1} = j | Q_t = i) \qquad (12) \\
&= \sum_{j=1}^{2} A_{ij} b_j^{t+1} w_j^{t+1},
\end{aligned}
$$

for $t = N-1, N-2, \ldots, 1$, starting with $w_i^N = 1$. An immediate relation between forward and backward variables is given by $p_\theta(\mathbf{X}^{(N)}, Q_t = i) = v_i^t w_i^t$.

2.5.1.3. Posterior regime probabilities

Using the forward and backward variables, we may calculate the posterior probability $\gamma_i^t = \mathbb{P}(Q_t = i | \mathbf{X}^{(N)})$ of being in state i at time t given the entire sequence $X_0, \ldots, X_N$ of observations.

$$
\gamma_i^t = \frac{p_\theta(\mathbf{X}^{(N)}, Q_t = i)}{p_\theta(\mathbf{X}^{(N)})} = \frac{p_\theta(\mathbf{X}^{(N)}, Q_t = i)}{\sum_{k=1}^{2} p_\theta(\mathbf{X}^{(N)}, Q_t = k)} = \frac{v_i^t w_i^t}{\sum_{k=1}^{2} v_k^t w_k^t}. \qquad (13)
$$

Mark that γ_i^t is the conditional expectation of $S_{t,i}$ given the data $\mathbf{X}^{(N)}$ as the coordinates of the state vector S_t are 0-1-variables, i.e.,

$$
\mathbb{E}\{S_{t,i} | \mathbf{X}^{(N)}\} = \mathbb{P}(S_{t,i} = 1 | \mathbf{X}^{(N)}) = \gamma_i^t. \qquad (14)
$$

We also need the joint posterior probability $\xi_{ij}^{t,t+1} = \mathbb{P}(Q_t = i, Q_{t+1} = j | \mathbf{X}^{(N)})$ of the switching variables at time t and $t+1$ for which we have

$$
\xi_{ij}^{t,t+1} = \frac{p_\theta(\mathbf{X}^{(N)}, Q_t = i, Q_{t+1} = j)}{p_\theta(\mathbf{X}^{(N)})} = \frac{A_{i,j} v_i^t b_j^{t+1} w_j^{t+1}}{\sum_{k=1}^{2} v_k^t w_k^t}, \qquad (15)
$$

since, using (A1), (A2),

$$
\begin{aligned}
& p_\theta(\mathbf{X}^{(N)}, Q_t = i, Q_{t+1} = j) \\
&= p_\theta(X_{t+2}, \ldots, X_N | \mathbf{X}^{(t+1)}, Q_t = i, Q_{t+1} = j) p_\theta(\mathbf{X}^{(t+1)}, Q_t = i, Q_{t+1} = j) \\
&= p_\theta(X_{t+2}, \ldots, X_N | X_{t+1}, Q_{t+1} = j) p_\theta(\mathbf{X}^{(t+1)}, Q_t = i, Q_{t+1} = j) \\
&= w_j^{t+1} p(X_{t+1} | \mathbf{X}^{(t)}, Q_t = i, Q_{t+1} = j) p_\theta(\mathbf{X}^{(t)}, Q_t = i, Q_{t+1} = j) \\
&= w_j^{t+1} p_\theta(X_{t+1} | X_t, Q_{t+1} = j) \mathbb{P}(Q_{t+1} = j | Q_t = i, \mathbf{X}^{(t)}) p_\theta(\mathbf{X}^{(t)}, Q_t = i) \\
&= A_{i,j}\, v_i^t\, b_j^{t+1}\, w_j^{t+1}.
\end{aligned}
$$

In an online setting, it is sometimes more interesting, in particular for forecasting purposes, to get the conditional distribution of the switching variable Q_t at time t given the

observations up to time $t-1$ only instead of the whole sample. By similar arguments, we have, e.g., instead of (14)

$$\mathbb{E}\{S_{t,i}|\mathbf{X}^{(t-1)}\} = \mathbb{P}(Q_t = i|\mathbf{X}^{(t-1)}) = \frac{\sum_{i=1}^{2} v_i^{t-1} A_{i,k}}{\sum_{j=1}^{2}\sum_{i=1}^{2} v_i^{t-1} A_{i,j}}.$$

2.5.2. Maximization or M-step

In this section, we consider the state variables Q_t or, equivalently, $S_{t,i}, i = 1, 2$, to be known. Moreover, we set

$$Z_{ij}^{t,t+1} = S_{t,i} S_{t+1,j} = 1 \quad \text{iff } Q_t = i, Q_{t+1} = j.$$

In the final iteration scheme, $S_{t,i}, Z_{ij}^{t,t+1}$ will be replaced by estimates of their conditional expectations given the data, calculated during the E-step. We use our supposed knowledge of the hidden Markov chain to get estimates of the transition matrix A and the autoregressive parameters of the regimes by maximizing the complete log-likelihood

$$\ell_c\left(\theta|\mathbf{X}^{(N)}, \mathbf{Q}^{(N)}\right) = \log \pi_{Q_0} + \sum_{t=1}^{N} \log A_{Q_{t-1},Q_t} + \ell_c^{AR}\left(\theta|\mathbf{X}^{(N)}, \mathbf{Q}^{(N)}\right).$$

Neglecting the terms involving Q_0, which does not make a big difference for large enough N, we get from setting the derivative of ℓ_c with respect to A_{ii} to 0 and recalling $A_{i1} + A_{i2} = 1$

$$\hat{A}_{ii} = \frac{\frac{1}{N-1}\sum_{t=1}^{N-1} Z_{ii}^{t,t+1}}{\frac{1}{N}\sum_{t=1}^{N} S_{t,i}}, \quad i = 1, 2, \tag{16}$$

i.e., the ratio of the relative frequency of transitions from state i to state i and the relative frequency of visits in state i. Correspondingly, we estimate the stationary probabilities $\pi_i = \mathbb{P}(Q_t = i)$ of the Markov chain by

$$\hat{\pi}_i = \frac{1}{N}\sum_{t=1}^{N} S_{t,i}, \tag{17}$$

i.e., the relative number of visits in regime i.

To maximize $\ell_c^{AR}(\theta|\mathbf{X}^{(N)}, \mathbf{Q}^{(N)})$ with respect to $\alpha_1, \alpha_2, \sigma_1^2, \sigma_2^2$, we have to minimize, compare (4),

$$G(\theta) = \frac{1}{2}\sum_{t=1}^{N}\sum_{k=1}^{2} S_{tk}\left(\log \sigma_k^2 + \frac{(X_t - \alpha_k X_{t-1})^2}{\sigma_k^2}\right). \tag{18}$$

Mark that $G(\theta)$ does not depend on A_{11}, A_{22}. From solving $\frac{\partial}{\partial \alpha_k} G(\theta) = 0$, we get

$$\hat{\alpha}_k = \frac{\sum_{t=1}^{N} S_{t,k} X_{t-1} X_t}{\sum_{t=1}^{N} S_{t,k} X_{t-1}^2}, \tag{19}$$

i.e., the first sample autocorrelation of the kth regime. Similarly, solving $\frac{\partial}{\partial \sigma_k^2} G(\theta) = 0$ results in

$$\hat{\sigma}_k^2 = \frac{\sum_{t=1}^N S_{t,k}(X_t - \hat{\alpha}_k X_{t-1})^2}{\sum_{t=1}^N S_{t,k}}, \tag{20}$$

i.e., the usual residual variance estimate of the kth regime.

Mark that it is a special feature of Markov switching autoregressions that we get explicit formulas for the estimates of the regime parameters $\alpha_1, \alpha_2, \sigma_1^2, \sigma_2^2$. For more complex, e.g., nonlinear autoregressive, models, those estimates have to be calculated numerically.

2.5.3. *The combined algorithm*

The forward–backward procedure and the maximization of the likelihood are now iteratively combined to form the final EM algorithm. We start with some initial value $\hat{\theta}(0) = (\hat{\alpha}_1(0), \hat{\alpha}_2(0), \hat{\sigma}_1^2(0), \hat{\sigma}_2^2(0), \hat{A}_{11}(0), \hat{A}_{22}(0))^T$ for the parameter θ, and we set the iteration index $n = 0$. Then, we iterate between the E-step of Section 2.5.1, where the unknown parameters are replaced by their current estimates, and the M-step of Section 2.5.2, where the hidden switching variables are replaced by their conditional expectations given the data:

2.5.3.1. *E-step*

Assume that the model parameters are given by $\hat{\theta}(n)$. Compute for $t = 1, \ldots, N$, the forward variables $v_k^t(n)$ and the backward variables $w_k^t(n)$ from (10), (12). Calculate the auxiliary variables $\gamma_i^t(n)$ and $\xi_{ij}^{t,t+1}(n)$ from (13), (15).

2.5.3.2. *M-step*

Calculate the updated estimate $\hat{A}(n+1), \hat{\pi}(n+1)$ from (16), (17) replacing $S_{t,i}, Z_{ij}^{t,t+1}$ by $\gamma_i^t(n), \xi_{ij}^{t,t+1}(n)$ from the E-step.

Calculate the autoregressive parameters $\hat{\alpha}_1(n+1), \hat{\alpha}_2(n+1), \hat{\sigma}_1^2(n+1), \hat{\sigma}_2^2(n+1)$ from (19), (20), where, again, the $S_{t,k}$ are replaced by the current estimates $\gamma_k^t(n)$ of their conditional expectations given the data.

Combine the new estimates $\hat{\alpha}_i(n+1), \hat{\sigma}_i^2(n+1), \hat{A}_{ii}(n+1), i = 1, 2$, to get the updated parameter vector estimate $\hat{\theta}(n+1)$.

The iteration is continued for $m = 0, 1, 2, \ldots$ until a stopping criterion is satisfied.

For the simulated data of Fig. 1 with

$$\alpha_1 = 0.9, \alpha_2 = -0.9, \sigma_1 = 1, \sigma_2 = 0.5, A_{11} = 0.8, A_{22} = 0.9,$$

the EM algorithm stabilized rather soon. Already 10 iterations resulted in the following estimates that did not change further up to the fourth decimal by continuing the iteration further:

$$\hat{\alpha}_1 = 0.9550, \hat{\alpha}_2 = -0.8872, \hat{\sigma}_1 = 1.0018,$$

$$\hat{\sigma}_2 = 0.4297, \hat{A}_{11} = 0.8352, \hat{A}_{22} = 0.9393.$$

The initial values in $\hat{\theta}(0)$ have been chosen by a random number generator.

2.6. The Viterbi algorithm

The EM algorithm does not only provide a numerical method for calculating the maximum likelihood estimates of the model parameters, but simultaneously results in estimates $\gamma_i^t(n)$ of the posterior state probabilities $\mathbb{P}(Q_t = i|\mathbf{X}^{(N)})$ given the data. In the case of independent switching variables Q_t discussed by Wong and Li (2000) or Franke et al. (2011), those immediately provide a MAP estimate for the regime at time t:

$$\widehat{Q}_t = 1, \quad \text{if } \gamma_1^t(n) > \gamma_2^t(n), \quad \widehat{Q}_t = 2, \quad \text{otherwise.}$$

In general, however, one has to take into account the dependencies between the Q_t and to look simultaneously for the whole most plausible sequence $\widehat{Q}_1, \ldots, \widehat{Q}_N$ of regimes given the data. The Viterbi algorithm, compare, e.g., Rabiner and Jiang (1986), provides a solution to that problem. We assume again that the model parameter θ is given; in practice, it has to be replaced by its estimate from the EM algorithm. We define

$$\delta_i^t = \max_{q_1,\ldots,q_{t-1}} \log p_\theta(\mathbf{X}^{(t)}, q_1, q_2, \ldots, q_{t-1}, Q_t = i),$$

i.e., δ_i^t is the highest complete log-likelihood along a single path of the Markov chain up to time t, which ends in state $Q_t = i$. As from (A1), (A2)

$$\begin{aligned} &p_\theta(\mathbf{X}^{(t+1)}, q_1, q_2, \ldots, q_t, Q_{t+1} = j) \\ &= \mathbb{P}(Q_{t+1} = j|Q_t = q_t)\, p_\theta(X_{t+1}|X_t, Q_{t+1} = j)\, p_\theta(\mathbf{X}^{(t)}, q_1, \ldots, q_t), \end{aligned}$$

we get the recursion

$$\delta_j^{t+1} = \max_i(\delta_i^t + \log A_{ij}) + \log b_j^{t+1},$$

compare (11). We also need the auxiliary variables I_j^t that correspond to that index i for which the maximum in the previous relation is attained. Using the initial values

$$\delta_j^1 = \log \pi_j b_j^1, \quad I_j^1 = 0, \quad j = 1, 2,$$

the recursion can be written in the form

$$\begin{aligned} I_j^t &= \arg\max_{i=1,2} \left(\delta_i^{t-1} + \log A_{ij}\right), \quad 2 \le t \le N, \quad j = 1, 2, \\ \delta_j^t &= \delta_{I_j^t}^{t-1} + \log A_{I_j^t, j} + \log b_j^t, \quad 2 \le t \le N, \quad j = 1, 2. \end{aligned}$$

We finally get the most plausible terminal state

$$\widehat{Q}_N = \arg\max_{i=1,2} \left(\delta_i^N\right)$$

To retrieve the whole state sequence, we have to backtrack the most plausible single regimes from time N to time 1:

$$\widehat{Q}_t = I_{\widehat{Q}_{t+1}}^{t+1}, \quad t = N-1, N-2, \ldots, 1.$$

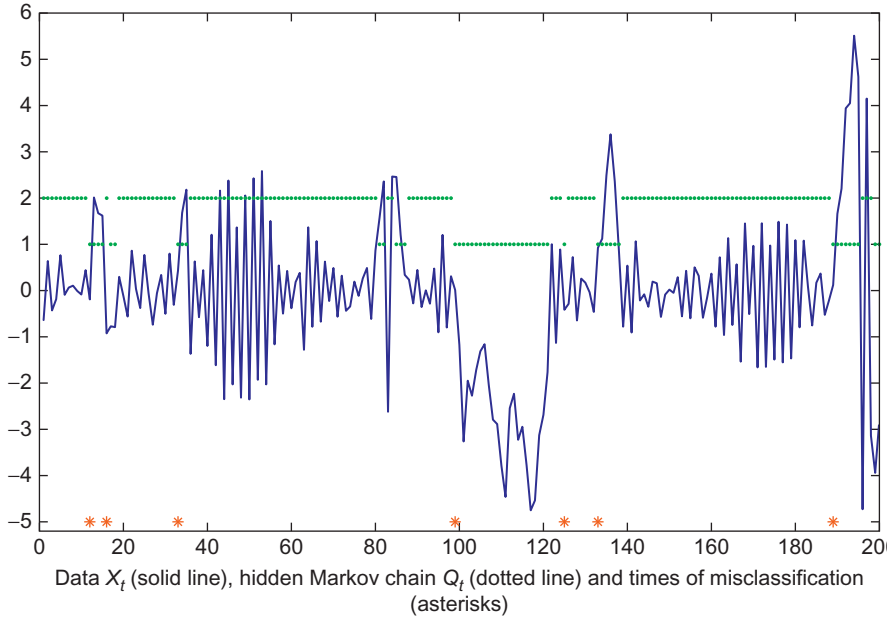

Fig. 3. Data X_t from Fig. 1 together with the hidden Markov chain Q_t.

For the data of Fig. 1, the Viterbi algorithm resulted in an almost perfect reconstruction of the hidden sequence Q_t. We have $\widehat{Q}_t = Q_t$ for 193 out of 200 observations. Figure 3 shows the data (solid line) and the hidden Markov chain Q_t (dotted line). The seven time instants where the estimated states did not coincide with the true ones are marked with an asterisks at the bottom of the figure.

3. Other Markov switching time series models

Most of the literature mentioned above allows for multivariate versions of the Markov switching autoregressive model (2). Such regime-switching vector autoregressions and their applications in economics are discussed in particular in the monograph by Krolzig (1997). Francq and Zakoian (2001) have investigated Markov switching ARMA models and their properties like the existence of stationarity realizations. Markov switching autoregressions have been also extended to include exogenous observable time series variables. Such Markov switching ARX processes are discussed in Section 12.3 of Frühwirth-Schnatter (2006), where they are called Markov switching dynamic regression models.

Yao and Attali (2000) investigate the structure of general nonparametric Markov switching autoregressions, corresponding to (9) with an arbitrary function F that is not characterized by a finite-dimensional parameter θ. Franke et al. (2011) discuss nonparametric estimates for the autoregressive function based on smoothing kernels combined with an EM algorithm. Franke et al. (2010) have extended this setup to nonparametric

Markov switching AR–ARCH models which they call conditionally heteroscedastic mixtures of experts (CHARME). In the case of autoregressive and ARCH components of order 1, this model assumes the form

$$X_t = \sum_{k=1}^{K} S_{tk} \left(\alpha_k(X_{t-1}) + \sigma_k(X_{t-1})\epsilon_t\right).$$

Tadjuidje et al. (2005) investigate nonparametric estimates for the autoregressive and volatility functions $\alpha_k(x), \sigma_k(x), k = 1, \ldots, K$, based on fitting feed-forward neural networks and applies them to a portfolio management problem. The identifiability problem for that particular type of nonlinear Markov switching autoregressions is treated by Stockis et al. (2008).

Parametric Markov switching models of the ARCH/GARCH type have been extensively discussed in the literature. A simple example is the Markov switching ARCH (MS-ARCH) process of order 1 with K regimes which is used as a model for asset returns X_t,

$$X_t = \sigma_t\epsilon_t; \quad \sigma_{t+1}^2 = \sum_{t=1}^{K} S_{tk} \left(\omega_k + \alpha_k X_t^2\right).$$

This kind of model and extensions to higher order ARCH and to Markov switching GARCH processes have been considered by Cai (1994), Hamilton and Susmel (1994), Wong and Li (2001), Francq et al. (2001), Kaufmann and Frühwirth-Schnatter (2002), and Francq and Zakoian (2005). Maximum likelihood estimation in the case of the straightforward generalization to Markov switching GARCH processes encounters inherent difficulties as the dependency on the path of the hidden Markov chain leads to a number of terms in the likelihood function which increases exponentially with sample size. To handle this problem, Haas et al. (2004) and Lanne and Saikkonen (2003) have proposed modified approaches to Markov switching GARCH allowing for feasible parameter estimates as well as volatility forecasts.

An interesting direction for future research is the study of models like that of Tadjuidje et al. (2009), which combine the ideas of Markov switching and observation-based switching. Here, a Markov switching time series model is considered, where the actual transition probabilities of the Markov chain from time t to time $t + 1$ depend on the observed data up to time t. In the notation of section 2, the transition matrix A would be a parametric function of past values $X_s, s \leq t$.

4. Markov switching in continuous time

Hidden Markov and Markov switching models have recently become popular in mathematical finance where mainly stochastic processes in continuous time are of interest. Rydén et al. (1998) have pointed out early that discrete-time hidden Markov models can reproduce the stylized facts of asset prices (listed, e.g., in Section 1.2 of Franses and van Dijk (2000)), and may therefore be used for modeling financial data. If we start from the classical Black–Scholes model, we may introduce regime switching by letting

the drift and volatility parameters μ, σ depend on the state of a hidden Markov process. We get for the price process Y_t and the returns R_t

$$dY_t = Y_t\, dR_t, \quad dR_t = \mu_{Q_t} + \sigma_{Q_t} dW_t,\ t \geq 0, \tag{21}$$

where $\{W_t\}$ is a standard Wiener process with respect to a filtration $\mathcal{F} = \{\mathcal{F}_t, t \geq 0\}$, and Q_t is a $\mathcal{F}$-adapted Markov process in continuous time assuming only finitely many values $1, \ldots, K$. The distribution of $\{Q_t\}$ is determined by a generator matrix G representing the rates and the transition probabilities for leaving the current regime. Elliott and Wu (2005) have considered the extension of (21) to jump-diffusion processes. Xi (2008) has investigated the more general model

$$dY_t = \mu_{Q_t}(Y_t)dt + \sigma_{Q_t}(Y_t)dW_t,\ t \geq 0,$$

with arbitrary functions μ_k, σ_k and given conditions for geometric ergodicity.

There is a specific difference between (21) and the analogous hidden Markov model $X_t = \mu_{Q_t} + \sigma_{Q_t}\epsilon_t$ in discrete time. If σ is not constant, then, in continuous time the Markov process Q_t can be reconstructed theoretically from the observations R_t as $\sigma^2_{Q_t}$ is known from the derivative of the quadratic variation of R_t. Therefore, this model is usually called a Markov switching instead of a hidden Markov model in the literature in contrast to our terminology in the discrete-time case.

The main ideas for developing estimation and filtering algorithms in continuous time are similar to those described above for discrete time. Hahn et al. (2009b) give a review of the recent literature. The interplay between discretized filters based on continuous-time models and the corresponding filters for discrete-time data is discussed by James et al. (1996). If, as frequently is the case in practice, only discrete observations of a continuous-time process are available, specific problems show up, e.g., there is not necessarily a corresponding generator matrix of a continuous-time Markov process given an arbitrary transition matrix of a discrete-time Markov chain. If we estimate the latter based on discrete data, then it is not clear how to get an estimate of the distribution of the underlying continuous-time process. To fill this gap, Hahn et al. (2009a,b) and Hahn and Sass (2009) develop appropriate numerical algorithms based on Markov chain Monte Carlo ideas that, however, are computationally quite expensive and require a considerable amount of data.

Erlwein et al. apply such continuous-time hidden Markov models to a variety of financial data like electricity spot prices (Erlwein et al., 2010) or short term interest rates (Erlwein et al., 2009). Applications to portfolio management problems haven been discussed recently by Bäuerle and Rieder (2004), Erlwein et al. (2009), Hahn et al. (2007), and Sass and Haussmann (2004).

Acknowledgments

The work was supported by the Center for Mathematical and Computational Modelling $(CM)^2$ funded by the state of Rhineland-Palatinate.

References

Bäuerle, N., Rieder, U., 2004. Portfolio-optimization with Markov-modulated stock prices and interest rates. IEEE Trans. Autom. Control 49, 442–447.

Baum, L.E., Petrie, T., 1966. Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Statist. 37, 1554–1563.

Baum, L.E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Statist. 41, 164–171.

Bauwens, L., Rombouts, J., 2010. On marginal likelihood computation in change-point models. Comp. Stat. Data Anal. doi:10.1016/j.csda.2010.06.025 to appear.

Brockwell, P., Davis, R., 1991. Time Series: Theory and Methods. Springer Verlag, Berlin, Heidelberg, New York.

Cai, J., 1994. A Markov model of switching-regime ARCH. J. Bus. Econ. Statist. 12, 309–316.

Cappé, O., Moulines, E., Rydén, T., 2005. Inference in Hidden Markov Models. Springer-Verlag, Berlin, Heidelberg, New York.

Carvalho, A.X., Tanner, M.A., 2005. Mixtures-of-experts of autoregressive time series: asymptotic normality and model specification. IEEE Trans. Neural Netw. 16, 39–56.

Chopin, N., 2007. Inference and model choice for sequentially ordered hidden Markov models. J. R. Stat. Soc. B 69, 269–284.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood for incomplete data via the em algorithm. J. R. Stat. Soc. B 39, 1–38.

Douc, R., Moulines, E., Rydén, T., 2004. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime Ann. Statist. 32, 2254–2304.

Elliott, R.J., Wu, P., 2005. Hidden Markov model filtering for jump diffusions. Stoch. Anal. Appl. 23, 153–163.

Erlwein, C., Benth, F.E., Mamon, R., 2010. HMM filtering and parameter estimation of an electricity spot price model. Energy Econ. doi: 10.1016/j.eneco.2010.01.005.

Erlwein, C., Mamon, R., 2009. An online estimation scheme for a Hull-White model with HMM-driven parameters. Stat. Methods Appl. 18, 87–107.

Erlwein, C., Mamon, R., Davison, M., 2009. An examination of HMM-based investment strategies for asset allocation. Appl. Stoch. Models Bus. Ind. DOI: 10.1002/asmb.802.

Francq, C., Roussignol, M., 1998. Ergodicity of autoregressive processes with Markov switching and consistency of the MLE. Statistics 32, 151–173.

Francq, C., Roussignol, M., Zakoian, J.M., 2001. Conditional heteroskedasticity driven by hidden Markov chains. J. Time Ser Anal. 22, 197–220.

Francq, C., Zakoian, J.M., 2001. Stationarity of multivariate Markov switching ARMA models. J. Econom. 102, 339–364.

Francq, C., Zakoian, J.M., 2005. The L^2 structures of standard and switching-regime GARCH models. Stoch. Proc. Appl. 115, 1557–1582.

Franke, J., Stockis, J.-P., Kamgaing, J.T., 2010. On geometric ergodicity of CHARME models. J. Time Ser. Anal. 31, 141–152.

Franke, J., Stockis, J.-P., Kamgaing, J.T., Li, W.K., 2011. Mixtures of nonparametric autoregressions. J. Nonpar. Stat. 23, 287–303.

Franses, P.H., van Dijk, D., 2000. Non-linear Time Series Models in Empirical Finance. Cambridge University Press, Cambridge.

Frühwirth-Schnatter, S., 2001. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models J. Amer. Statist. Ass. 96, 194–209.

Frühwirth-Schnatter, S., 2004. Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. Econom. J. 7, 143–167.

Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer-Verlag, Berlin, Heidelberg, New York.

Guidolin, M., Timmermann, A., 2007. Asset allocation under multivariate regime switching. J. Econ. Dyn. Control, 31, 3503–3544.

Haas, M., Mittnik, S., Paolella, M.S., 2004. Autoregressive conditional heteroskedasticity and changes in regime. J. Financial Econom., 2, 493–530.

Hahn, M., Putschögl, W., Sass J., 2007. Portfolio optimization with non-constant volatility and partial information. Braz. J. Probab. Stat. 21, 27–61.

Hahn, M., Frühwirth-Schnatter, S., Sass J., 2009. Estimating continuous-time Markov processes based on merged time series. Adv. Stat. Anal. 93, 403–425.

Hahn, M., Frühwirth-Schnatter, S., Sass J., 2009. Markov chain Monte Carlo methods for parameter estimation in multidimensional continuous time Markov switching models. J. Financial Econom. 8, 88–121.

Hahn, M., Sass J., 2009. Parameter estimation in continuous time Markov switching models – A semi-continuous Markov chain Monte Carlo approach. Bayesian Anal. 4, 63–84.

Hamilton J.D., 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57, 357–384.

Hamilton J.D., 1990. Analysis of time series subject to changes in regime. J. Econom. 45, 39–70.

Hamilton J.D., Susmel R., 1994. Autoregressive conditional heteroskedasticity and changes in regime. J. Econom. 64, 307–333.

Holst U., Lindgren G., Holst J., Thuvesholmen M., 1994. Recursive estimation in switching autoregressions with Markov regime. J. Time Ser. Anal. 15, 489–506.

James M.R., Krishnamurthy V., Le Gland, F., 1996. Time discretization of continuous-time filters and smoothers for HMM parameter estimation. IEEE Trans. Inf. Theory 42, 593–605.

Jiang, Tanner M.A., 1999. On the identifiability of mixture-of-experts. Neural Netw. 12, 197–220.

Kaufmann S., Frühwirth-Schnatter, S., 2002. Bayesian analysis of switching ARCH models. J. Time Ser. Anal. 23, 425–458.

Krolzig H.-M., 1997. Markov Switching Vector Autoregressions. Modelling, Statistical Inference and Application to Business Cycle Analysis. Lecture Notes in Econom. and Math. Systems. 454, Springer Verlag, Berlin, Heidelberg, New York.

Lange T., Rahbek, A., 2009. An introduction to regime switching time series models. In: Andersen, T.G., Davis, R.A., Kreiß J.-P., Mikosch, T. (Eds.), Handbook of Financial Time Series. Springer-Verlag, Berlin, Heidelberg, New York.

Lanne, M., Saikkonen, P., 2003. Modeling the US short-term interest rate by mixture autoregressive processes. J. Financial Econom 1, 96–125.

Liehr, S., Pawelzik, K., Kohlmorgen, J., Müller, K.-R., 1999. Hidden Markov mixtures of experts with an application to EEG recordings from sleep. Theory Biosci. 118, 246–260.

Liehr, S., Pawelzik, K., Kohlmorgen, J., Lemm, S., Müller, K.-R., 1999. Hidden Markov mixtures of experts for prediction of non-stationary dynamics. Neural Netw. Signal Process. Proc. 1999 IEEE Signal Process. Society Workshop 9, 195–204.

Lindgren, G., 1978. Markov regime models for mixed distributions and switching regressions. Scand. J. Stat. 5, 81–91.

MacDonald, I,L., Zucchini, W., 1997. Hidden Markov and Other Models for Discrete-Valued Time Series. Chapman and Hall, London.

MacKay, R.J., 2002. Estimating the order of a hidden Markov model. Can. J. Stat. 30, 573–589.

McCulloch, R.E., Tsay, R.S., 1994. Statistical analysis of economic time series via Markov switching models. J. Time Ser Anal. 15, 523–539.

Müller, K.-R., Kohlmorgen, J., Rittweger, J., Pawelzik, K., 1995. Analysing physiological data on wake-sleep state transition with competing predictors. NOLTA 95: Las Vegas Symposium on Nonlinear Theory and its Applications, IEICE, Tokyo , 223–226.

Peng, F., Jacobs, R.A., Tanner, M.A., 1996. Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. J. Am. Stat. Ass. 91, 953–960.

Pinson, P., Christensen, L.E.A., Madsen, H., Sørensen, P.E., Donovan, M.H., Jensen, L.E., 2008. Regime-switching modelling of the fluctuations of offshore wind generation. J. Wind Eng. Ind. Aerodynamics 96, 2327–2347.

Psaradakis, Z., Spagnolo, N., 2003. On the determination of the of the number of regimes in Markov-switching autoregressive models. J. Time Ser. Anal. 24, 237–252.

Psaradakis, Z., Spagnolo, N., 2006. Joint determination of the state dimension and autoregressive order for models with Markov regime switching. J. Time Ser. Anal. 27, 753–766.

Rabiner, L., Jiang, B., 1986. An introduction to hidden Markov models. IEEE ASSP Magazine, 3, 4–16.

Rydén, T., 1995. Estimating the order of hidden Markov models. Statistics 26, 345–354.

Rydén, T., 2008. EM versus Markov chain Monte Carlo for estimation of hidden Markov models: A computational perspective. Bayesian Anal. 08, 659–688.

Rydén, T., Teräsvirta, T., Asbrink, S., 1998. Stylized facts of daily return series and the hidden Markov model. J. Appl. Econ. 13, 217–244.

Sass, J., Haussmann, U.G., 2004. Portfolio optimization under partial information: Stochastic volatility in a hidden Markov model. In: Ahr, D., Fahrion, R., Oswald, M., Reinelt, G. (Eds.), Operations Research Proceedings 2000. Springer Verlag, Berlin, Heidelberg, New York.

Stockis, J.P., Tadjuidje Kamgaing, J., Franke, J., 2008. A note on the identifiability of the conditional expectation for mixtures of neural networks. Stat. Probab. Lett 78, 739–742.

Tadjuidje Kamgaing, J., Ombao, H., Davis, R.A., 2005. Competing neural networks as models for nonstationary financial time series. PhD Thesis, Dept. of Mathematics, University of Kaiserslautern.

Tadjuidje Kamgaing, J., Ombao, H., Davis, R.A., 2009. Autoregressive processes with data-driven regime switching. J. Time Ser. Anal. 30, 505–533.

Timmermann, A., 2000. Moments of Markov switching models. J. Econom. 96, 75–11.

Tong, H., 1990. Non-linear Time Series. Oxford University Press, Oxford.

Wong, C.S., Li, W.K., 2000. On a mixture autoregressive model. J. R. Stat. Soc. B 62, 95–115.

Wong, C.S., Li, W.K., 2001. On a mixture autoregressive conditional heteroscedastic model. J. Am. Stat. Assoc. 96, 982–995.

Wu, C.F.J., 1983. On the convergence properties of the EM algorithm. Ann. Stat. 11, 95–103.

Xi, F.B., 2008. Feller property, exponential ergodicity of diffusion processes with state-dependent switching. Sci. China Ser. A 51, 329–342.

Yao, F., Attali, J., 2000. On stability of nonlinear AR processes with Markov switching. Adv. Appl. Probab. 32, 394–407.

Zhang, J., Stine, R.A., 2001. Autocovariance structure of Markov regime switching models and model selection. J. Time Ser. Anal. 22, 107–124.

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00006-5

A Review of Robust Estimation under Conditional Heteroscedasticity

Kanchan Mukherjee

Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK

Abstract

In this chapter, we discuss estimation of parameters for heteroscedastic models. In particular, we discuss the class of M-estimators for the parameters of the symmetric as well as asymmetric heteroscedasticity and the classes of rank and M-estimators of the parameters associated with the conditional mean function of the autoregressive models. We investigated robustness properties of the proposed estimators through extensive simulation and financial data analysis.

Keywords: Heteroscedastic models, Rank and M-estimation, VaR.

1. Introduction

Observed financial series frequently appear uncorrelated, yet they exhibit volatility clustering. Volatility clustering is the tendency of observations relatively large (small) in absolute values to be followed by other large (small) observations. Nonlinear models with time-dependent conditional variance are often used to describe time series with this feature. Engle (1982) introduced the autoregressive conditional heteroscedastic (ARCH) model to describe the inflation rate. ARCH models are used to represent the volatility, i.e., the strong dependence of the instantaneous variability of a time series on its own past. Since its introduction, there have been huge developments on the theory and application of this model and its various generalizations to economics and financial data sets. In this chapter, we discuss various estimation procedures associated with such models and their applications.

A popular method for estimating the unknown parameters in such models is to use the Gaussian likelihood of the innovations and the resulting estimator is called the

quasi maximum likelihood estimator (QMLE). The QMLE is consistent and asymptotically normal when the innovation distribution has finite fourth moment. However, such stringent moment condition may not hold in many situations; an example is innovations with student-t distribution where the degrees of freedom is at most 4. To deal with such situations, we discuss robust estimation procedures for these models. See, for example, Koul and Mukherjee (2002), Peng and Yao (2003), Berkes and Horvath (2004), Mukherjee (2006, 2007, 2008), Iqbal and Mukherjee (2010), among others for previous study in this direction.

In the first part of this chapter, we discuss M-estimation methods associated with the conditional variance parameters of both symmetric and asymmetric heteroscedastic models. In particular, we consider the generalized ARCH (GARCH) model of Bollerslev (1986), which is useful for modeling symmetric volatility. Another important consideration in volatility modeling is that unexpected changes in the return have different effects on the conditional variance; an unexpected increase (good news) contributes less to the conditional variance in the model than an unforeseen fall (bad news). Glosten et al. (1993) proposed an asymmetric model, popularly known as the GJR model for this purpose, and we discuss the M-estimation methods for the GJR model as well.

Value-at-Risk (VaR) is a commonly used statistic for measuring potential risk in financial market. VaR is the conditional quantile of the return distribution; see the work done by Jorion (2000) for a general introduction and exposition of VaR. One of the important steps in the estimation of VaR involves obtaining an estimate of the instantaneous variability or the volatility of a financial time series. We consider robust measures of VaR using M-estimators of the GARCH and GJR parameters. The performance of the proposed VaR estimates is extensively studied for three important financial data sets (S&P500, FTSE100, NIKKEI225). Both in-sample and out-of-sample VaR estimates are evaluated. The accuracy of the proposed one-day-ahead VaR estimates is discussed using a number of M-test statistics of this chapter. The robustness of these VaR estimates to the functional forms of the assumed models is also addressed. See the work done by Iqbal and Mukherjee (in press) also for a recent study on this.

Although most of the existing methodological literature have focused on developing estimation procedures for the parameters associated with the conditional variability, the development of the estimation methods associated with the conditional mean component of a heteroscedastic problem is also important from the application point of view and this has been largely overlooked. For this, in the second part of this chapter, we discuss M-estimation and a rank-based robust procedure for estimating the mean parameter of a nonlinear autoregressive model with conditional heteroscedastic errors.

The results reviewed in this chapter are important from a number of different angles. Since QMLE is a member of the class of M-estimators, in many senses, the M-methods discussed here are applicable to most of the previous analysis using the QMLE. Moreover, since we use nonparametric setup for the error distributions and some of the robust estimators of the mean and the heteroscedastic parameters used for the VaR evaluation are consistent and asymptotically normal under minimal moment assumption such as merely finite second moment of the innovations, some M-estimators are expected to perform well for those financial data for which the use of the QMLE cannot be justified due to lack of fourth moment. In fact, our empirical study indicates that in most cases M-estimators such as Cauchy and B-estimator predict the VaR more accurately than the

frequently used QMLE. The robustness of the VaR estimates to the model misspecification is a desirable property for practitioners. Thus, this review strengthens the point of using robust estimators for fitting the heteroscedastic models and predicting VaR.

2. GARCH (p, q) and GJR $(1, 1)$ models

In the GARCH (p, q) model, where $p,\ q \geq 1$ are known integers, the following representation of the series $\{X_t; t \in \mathcal{Z}\}$ is assumed:

$$X_t = \sigma_t \epsilon_t, \tag{1}$$

where $\{\epsilon_t; t \in \mathcal{Z}\}$ are unobservable i.i.d. errors symmetric about zero and

$$\sigma_t^2 = \omega_0 + \sum_{i=1}^{p} \alpha_{0i} X_{t-i}^2 + \sum_{j=1}^{q} \beta_{0j} \sigma_{t-j}^2,\ t \in \mathcal{Z}, \tag{2}$$

with $\omega_0, \alpha_{0i},\ \beta_{0j} > 0, \forall\, i, j$. Bougerol and Picard (1992) discussed necessary and sufficient conditions for the existence of stationary solution to (1) and (2). Here, we are concerned with the problem of robust M-estimation of some function of the model parameter

$$\boldsymbol{\theta}_0 = [\omega_0, \alpha_{01}, \ldots, \alpha_{0p}, \beta_{01}, \ldots, \beta_{0q}]' \tag{3}$$

belonging to the parameter space Θ based on the observations $\{X_t; 1 \leq t \leq n\}$ and their applications.

Likelihood based on standardized normal distribution of $\{\epsilon_t\}$ is routinely used to estimate the parameters and the resulting estimator is called the QMLE. The asymptotic normality of the QMLE was established under the existence of unconditional error moments of order at least 4. Berkes et al. (2003) (abbreviated as BHK) derived many nice technical results on the GARCH model (1) and (2) and used them to derive the asymptotic normality of the QMLE.

Several studies on financial data however have suggested that the existence of fourth moment needed for the asymptotic normality of the QMLE is not tenable quite often in practice. Peng and Yao (2003) considered least absolute deviation (LAD)-type estimators of three different varieties and Berkes and Horvath (2004) and Mukherjee (2006) considered the pseudo-maximum likelihood estimator (PMLE) for the GARCH (p, q) model and ARCH (p) model. Berkes and Horvath (2004) derived asymptotic normality of some of the PMLEs under the existence of a fractional unconditional moment of the error distribution when the score function is three times differentiable over $(0, \infty)$. Their class of estimators include both LAD and QMLE as well as some other important score functions. However, the identifiability condition of the parameters to be estimated stipulates known value for the unconditional error (or function of error) moment such as $E(\epsilon^2) = 1$ or $E(|\epsilon|) = 1$ or $E\{|\epsilon|/(1 + |\epsilon|)\}$ is known; see the general condition (1.16) and displays (2.1)–(2.3) of specific examples in the study by Berkes and Horvath (2004). Clearly, such conditions are impossible to verify and hence are very undesirable. In this chapter, we discuss the asymptotics and applications of the PMLE, or more

generally, of the M-estimators, without making such assumptions. In particular, we note that in the GARCH model, an M-estimator based on a score function H consistently estimates

$$\boldsymbol{\theta}_{0H} = [c_H\omega_0, c_H\alpha_{01}, \ldots, c_H\alpha_{0p}, \beta_{01}, \ldots, \beta_{0q}]', \tag{4}$$

where c_H is a constant defined in (20) below, which depends on the score function H through the error distribution. In particular, an M-estimator can estimate $\boldsymbol{\theta}_0$ if and only if $c_H = 1$. Hence, using the QMLE, we can estimate $\boldsymbol{\theta}_0$ if and only if the error variance is unity, which is a standard assumption in the literature. See also the study by Fan et al. (2010) for similar findings.

In the GJR (1, 1) model, the following representation of the return series $\{X_t; t \in \mathcal{Z}\}$ is assumed.

$$X_t = \sigma_t\epsilon_t, \tag{5}$$

where $\{\epsilon_t; t \in \mathcal{Z}\}$ are unobservable i.i.d. errors symmetric about zero,

$$\sigma_t^2 = \omega_0 + \alpha_0 X_{t-1}^2 + \beta_0\sigma_{t-1}^2 + \gamma_0 D_{t-1}X_{t-1}^2, \quad D_{t-1} = I(X_{t-1} < 0) \tag{6}$$

and the unknown parameter is

$$\boldsymbol{\theta}_0 = [\omega_0, \alpha_0, \gamma_0, \beta_0]'. \tag{7}$$

Note that positive return contributes to the volatility through the factor α_0, whereas negative return increases the volatility through the factor $\alpha_0 + \gamma_0$. Here γ_0 is called the asymmetric parameter.

We assume that $\boldsymbol{\theta}_0$ is in the parameter space

$$\Theta = \{\boldsymbol{\theta} = [\omega, \alpha, \gamma, \beta]'; \omega, \alpha, \beta > 0, \alpha + \gamma \geq 0, \alpha + \beta + (\gamma/2) < 1\}.$$

Under these parameter constraints, model defined by (5) and (6) is strictly stationary.

Although the QMLE based on Gaussian likelihood is frequently used to estimate the parameters of the GJR model, it does not perform well unless finiteness of the fourth moment holds. Hence similar to the GARCH, we propose the class of robust M-estimators for the GJR model. We note that in the GJR model, an M-estimator based on a score function H consistently estimates

$$\boldsymbol{\theta}_{0H} = [c_H\omega_0, c_H\alpha_0, c_H\gamma_0, \beta_0]', \tag{8}$$

where, as before, c_H is a constant defined in (20).

2.1. M-estimators

In the sequel, for a function g, $\dot{g}$ and $\ddot{g}$ will denote the first and second derivatives, respectively, whenever they exist and ϵ will denote a random variable having same distribution as $\{\epsilon_t, \in \mathcal{Z}\}$.

Let $\psi : I\!R \to I\!R$ be an odd function that is differentiable in all but finite number of points. Let $\mathcal{D} \subset I\!R$ denote the set of points where ψ is differentiable and let $\bar{\mathcal{D}}$ denote its complement. Let $H(x) := x\psi(x)$, $x \in I\!R$. Note that $H(-x) = H(x)$, $\forall x$. The function H will be called the "score function" for the M-estimation in the scale model. Examples are as follows.

Example 1. Least absolute deviation (LAD) score: Let $\psi(x) = \text{sign}\,(x)$. Then $\mathcal{D}^c = \{0\}$ and $H(x) = |x|$.

Example 2. Huber's k-score: Let $\psi(x) = xI(|x| \le k) + k\,\text{sign}\,(x)I(|x| > k)$, where $k > 0$ is a known constant. Then $\mathcal{D}^c = \{-k, k\}$ and $H(x) = x^2 I(|x| \le k) + k|x|I(|x| > k)$.

Example 3. QMLE: Let $\psi(x) = x$. Then $H(x) = x^2$.

Example 4. Score function for the maximum likelihood estimation (MLE): Let $\psi(x) = -\dot{f}_0(x)/f_0(x)$, where f_0 is the true density of ϵ, assumed to be known. Then $H(x) = x\{-\dot{f}_0(x)/f_0(x)\}$.

Example 5. B-estimator: Let $\psi(x) = B\,\text{sign}(x)/(1 + |x|)$, where $B > 1$ is a known constant. Then $\mathcal{D}^c = \{0\}$ and $H(x) = B|x|/(1 + |x|)$.

Example 6. Cauchy estimator: Let $\psi(x) = 2x/(1 + x^2)$. Then $H(x) = 2x^2/(1 + x^2)$.

Example 7. Score function for the exponential pseudo-maximum likelihood estimation (EPMLE): Let $\psi(x) = a|x|^{b-1}\text{sign}(x)$, where $a > 0$ and $1 < b \le 2$ are known constants. Such score can be motivated from the class of densities considered by Nelson (1991) and Robinson and Zaffaroni (2006) to model the innovations of the exponential GARCH model. Here $\mathcal{D}^c = \{0\}$ and $H(x) = a|x|^b$.

Next we define M-estimators. Recall that in the location model, an M-estimator is defined as solution to certain system of equations involving residual functions and we follow the same approach. Since $\epsilon_t = X_t/\sigma_t$, to define residual functions, we first discuss the concept of variance function related to the denominator of the residual as follows. We discuss this for the GARCH and GJR seperately.

For the GARCH model, assume that for some $\kappa > 0$,

$$E[|\epsilon|^\kappa] < \infty. \tag{9}$$

Then from Lemma 2.3 and Theorem 1 of BHK, σ_t^2 of (2) has the following unique almost sure representation:

$$\sigma_t^2 = c_0 + \sum_{j=1}^{\infty} c_j X_{t-j}^2, \quad t \in \mathcal{Z}, \tag{10}$$

where $\{c_j; j \ge 0\}$ are defined in (7) through (9) of BHK and in (12) below.

Define the variance function on the parameter space Θ by

$$v_t(\boldsymbol{\theta}) = c_0(\boldsymbol{\theta}) + \sum_{j=1}^{\infty} c_j(\boldsymbol{\theta}) X_{t-j}^2, \quad \boldsymbol{\theta} \in \Theta, t \in \mathcal{Z}, \tag{11}$$

where the coefficients $\{c_j(\boldsymbol{\theta}); j \geq 0\}$ are given in BHK (Section 3 and display (3.1)) with the property

$$c_j(\boldsymbol{\theta}_0) = c_j, \quad \forall j \geq 0.$$

Hence, from (10), the variance functions satisfy

$$\sigma_t = v_t^{1/2}(\boldsymbol{\theta}_0), \quad t \in \mathcal{Z}.$$

An example of (11) for the GARCH (1, 1) model with $\boldsymbol{\theta} = (\omega, \alpha, \beta)'$ is

$$c_0(\omega, \alpha, \beta) = \omega/(1-\beta),\ c_j(\omega, \alpha, \beta) = \alpha\beta^{j-1}, \quad j \geq 1. \tag{12}$$

For the GJR model, by recursive substitution from (6),

$$\begin{aligned}
\sigma_t^2 &= \omega_0 + \alpha_0 X_{t-1}^2 + \gamma_0 D_{t-1} X_{t-1}^2 \\
&\quad + \beta_0\{\omega_0 + \alpha_0 X_{t-2}^2 + \gamma_0 D_{t-2} X_{t-2}^2 + \beta_0 \sigma_{t-2}^2\} \\
&= \omega_0(1+\beta_0) + \alpha_0(X_{t-1}^2 + \beta_0 X_{t-2}^2) + \gamma_0(D_{t-1}X_{t-1}^2 + \beta_0 D_{t-2} X_{t-2}^2) \\
&\quad + \beta_0^2\{\omega_0 + \alpha_0 X_{t-3}^2 + \gamma_0 D_{t-3} X_{t-3}^2 + \beta_0 \sigma_{t-3}^2\} \\
&= \omega_0(1+\beta_0+\beta_0^2) + \alpha_0(X_{t-1}^2 + \beta_0 X_{t-2}^2 + \beta_0^2 X_{t-3}^2) \\
&\quad + \gamma_0(D_{t-1}X_{t-1}^2 + \beta_0 D_{t-2}X_{t-2}^2 + \beta_0^2 D_{t-3} X_{t-3}^2) + \beta_0^3 \sigma_{t-3}^2 \\
&= \frac{\omega_0}{(1-\beta_0)} + \alpha_0 \sum_{j=1}^{\infty} \beta_0^{j-1} X_{t-j}^2 + \gamma_0 \sum_{j=1}^{\infty} \beta_0^{j-1} D_{t-j} X_{t-j}^2.
\end{aligned} \tag{13}$$

Hence, for $\boldsymbol{\theta} \in \Theta$, define the variance function

$$v_t(\boldsymbol{\theta}) = \frac{\omega}{(1-\beta)} + \alpha \sum_{j=1}^{\infty} \beta^{j-1} X_{t-j}^2 + \gamma \sum_{j=1}^{\infty} D_{t-j} \beta^{j-1} X_{t-j}^2 \tag{14}$$

and note that

$$\sigma_t = v_t^{1/2}(\boldsymbol{\theta}_0), \quad t \in \mathcal{Z}.$$

Therefore, (1) and (5) can be rewritten as

$$X_t = \{v_t(\boldsymbol{\theta}_0)\}^{1/2} \epsilon_t, \quad 1 \leq t \leq n. \tag{15}$$

Next consider observable approximations $\{\hat{v}_t(\boldsymbol{\theta})\}$ of the processes $\{v_t(\boldsymbol{\theta})\}$ of (11) and (14) defined by

$$\hat{v}_t(\boldsymbol{\theta}) = c_0(\boldsymbol{\theta}) + I(2 \leq t) \sum_{j=1}^{t-1} c_j(\boldsymbol{\theta}) X_{t-j}^2, \quad \boldsymbol{\theta} \in \Theta, \quad 1 \leq t \leq n,$$

and

$$\hat{v}_t(\boldsymbol{\theta}) = \frac{\omega}{(1-\beta)} + \alpha \sum_{j=1}^{t-1} \beta^{j-1} X_{t-j}^2 + \gamma \sum_{j=1}^{t-1} D_{t-j} \beta^{j-1} X_{t-j}^2, \ \boldsymbol{\theta} \in \Theta \ 1 \le t \le n,$$

for the GARCH and GJR models, respectively. Therefore, from (15), we define the residual functions as

$$X_t / \{\widehat{v}_t(\boldsymbol{\theta})\}^{1/2}, \quad 1 \le t \le n. \tag{16}$$

In (15), if f denotes the error density, then the conditional density of X_t given past will be $v_t^{-1/2}(\boldsymbol{\theta}_0) f\{v_t^{-1/2}(\boldsymbol{\theta}_0) X_t\}$, $1 \le t \le n$. Hence, motivated by the conditional likelihood, one can define a random quantity as a minimizer of the negative log-likelihood function $(1/n) \sum_{t=1}^{n} [(1/2) \log v_t(\boldsymbol{\theta}) - \log f\{X_t / v_t^{1/2}(\boldsymbol{\theta})\}]$, $\boldsymbol{\theta} \in \Theta$, or as a solution of its derivative function

$$\sum_{t=1}^{n} (1/2)[1 - H^*\{X_t / v_t^{1/2}(\boldsymbol{\theta})\}]\{\dot{v}_t(\boldsymbol{\theta}) / v_t(\boldsymbol{\theta})\} = \mathbf{0},$$

where $H^*(x) := x\{-\dot{f}(x)/f(x)\}$.

More generally, with a score function $H(x) := x\psi(x)$, we can then define $\boldsymbol{\theta}_n$ in (15) as a solution to the equation

$$\sum_{i=1}^{n} (1/2) \left\{ 1 - H\{X_t / v_t^{1/2}(\boldsymbol{\theta})\} \right\} \{\dot{v}_t(\boldsymbol{\theta}) / v_t(\boldsymbol{\theta})\} = \mathbf{0}. \tag{17}$$

Note, however, that $\boldsymbol{\theta}_n$s are noncomputable since $v_t(\boldsymbol{\theta})$s are nonobservable. Hence, replacing $v_t()$ by $\widehat{v}_t(\boldsymbol{\theta})$ in (17), an M-estimator $\hat{\boldsymbol{\theta}}_n$ for the respective model based on a score H is defined as a solution to

$$\sum_{i=1}^{n} (1/2) \left\{ 1 - H\{X_t / \widehat{v}_t^{1/2}(\boldsymbol{\theta})\} \right\} \{\dot{\widehat{v}}_t(\boldsymbol{\theta}) / \widehat{v}_t(\boldsymbol{\theta})\} = \mathbf{0}. \tag{18}$$

For $H(x) = x^2$ of Example 3, $\hat{\boldsymbol{\theta}}_n$ is the celebrated QMLE, whereas for $H(x) = |x|$ of Example 1, $\hat{\boldsymbol{\theta}}_n$ can be called the LAD estimator.

Based on an M-estimator, from (16), the residuals are defined as

$$\hat{\epsilon}_t = X_t / \{\widehat{v}_t(\hat{\boldsymbol{\theta}}_n)\}^{1/2}, \quad 1 \le t \le n. \tag{19}$$

2.2. *Asymptotic distribution of* $\hat{\boldsymbol{\theta}}_n$

Asymptotic distributions are derived under the following assumptions.

Model assumptions: Either (1) and (2) or (5) and (6) are valid. The parameter space Θ is a compact set and its interior Θ_0 contains both $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_{0H}$ of either (3) and (4) or (7) and (8), respectively. Moreover, (15) hold and $\{X_t\}$ is stationary and ergodic.

Conditions on the score function
Identifiability condition: Corresponding to the score function H, there exists a unique number $c_H > 0$ satisfying

$$E[H(\epsilon/c_H^{1/2})] = 1. \tag{20}$$

Moment conditions:

$$E[H(\epsilon/c_H^{1/2})]^2 < \infty \text{ and } 0 < E\{(\epsilon/c_H^{1/2})\dot{H}(\epsilon/c_H^{1/2})\} < \infty. \tag{21}$$

Smoothness conditions:
One can assume smoothness conditions of varying degree that are applicable to different score functions. One such (strong) assumption is that the score function is three times differentiable with bounded third derivative. It is possible to have weak smoothness conditions on H that are satisfied by all score functions of Examples 1–7.

To state the main result on $\hat{\boldsymbol{\theta}}_n$ of (18), define the score function factor

$$\sigma^2(H) := 4 \text{ var}\{H(\epsilon/c_H^{1/2})\}/[E\{(\epsilon/c_H^{1/2})\dot{H}(\epsilon/c_H^{1/2})\}]^2,$$

where $\text{var}\{H(\epsilon/c_H^{1/2})\}$ is assumed to be positive in the moment condition (21). Also, define

$$\boldsymbol{G} := E\{\dot{v}_1(\boldsymbol{\theta}_{0H})\dot{v}_1'(\boldsymbol{\theta}_{0H})/v_1^2(\boldsymbol{\theta}_{0H})\}.$$

THEOREM 1. *Suppose that the model assumptions, identifiability condition, moment conditions, and the smoothness conditions hold. Then*

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{0H}) \to N[0, \sigma^2(H)\boldsymbol{G}^{-1}]. \tag{22}$$

REMARK 1. The above result states that using the score function H, we can consistently estimate $\boldsymbol{\theta}_{0H}$. With $H(x) = x^2$, $c_H = E(\epsilon^2)$, and hence, using the QMLE, we can consistently estimate $[E(\epsilon^2)\omega_0, E(\epsilon^2)\alpha_{01}, \ldots, E(\epsilon^2)\alpha_{0p}, \beta_{01}, \ldots, \beta_{0q}]'$. Note that $E(\epsilon^2) = 1$ is a standard assumption in the literature, except in the study by Berkes and Horvath (2004) (for the GARCH), where known value of c_H is assumed for different H. Hence, when the error variance is unity, we can estimate $\boldsymbol{\theta}$ using the QMLE. We can estimate $\boldsymbol{\theta}$ using any other score function H whenever $c_H = 1$ for the corresponding error distribution. □

REMARK 2. Note that Theorem 1 is derived under weak moment assumptions on the error distribution. We imposed conditions on the score function H, which in most of the examples are translated to very mild moment assumptions on the error distribution. Also, the variance expressions (10) and (13) are crucial to define the variance functions. In particular, for Examples 5 and 6, only (9) is enough to have the variance expression for the GARCH, and hence, the asymptotic normality of the estimators where κ can be a fraction and need not even be of known value. □

The usefulness of the above estimators can be further demonstrated by considering a family of t density with $\nu > 0$ degrees of freedom where the density of the error random variable ϵ is proportional to

$$(1 + x^2/\nu)^{-(\nu+1)/2}. \tag{23}$$

Note that $E|\epsilon|^{\mu} < \infty$ for all $0 < \mu < \nu$, $E(\epsilon) = 0$ for $\nu > 1$, and $\mathrm{Var}(\epsilon) = \nu/(\nu - 2)$ for $\nu > 2$.

When $(2 < \nu \leq 4)$, for any b such that $2 < 2b < \nu$, $E|\epsilon|^{2b} < \infty$ but $E\epsilon^4 = \infty$. Therefore, the EPMLE satisfies (22), whereas the asymptotic normality of the QMLE does not hold.

The above class of estimators is useful for error distributions for which κ in (9) is possibly a fraction and even unknown. For illustration, suppose that the error density satisfies (23) for some unknown ν with $0 < \nu < 4$. Since (9) holds with $\kappa = \nu/2$, the estimator based on any known $\lambda > 1$ satisfies (22), whereas the asymptotic normality of the QMLE does not hold.

3. Data analysis for the GARCH and GJR models

We verify the asymptotic distributional result of Section 2 and compare relative performance of M-estimators based on different score functions using a simulation study. For comparison, we define the mean squared errors (MSEs) of an estimator for the GJR (1, 1) model as

$$E[\{(\hat{\omega} + \hat{\gamma})/\hat{\alpha} + \hat{\beta}\} - \{(\omega_0 + \gamma_0)/\alpha_0 + \beta_0)\}]^2.$$

From (22) and the definition of $\boldsymbol{\theta}_{0H}$, the ratio of $\hat{\omega} + \hat{\gamma}$ and $\hat{\alpha}$ is consistent to a quantity that is free from the underlying score function H used for the M-estimation, and hence, the above definition of MSE compares the relative performance of different M-estimators. When specialized to the GARCH (1, 1) model with $\gamma_0 = 0$, the corresponding MSE is defined as

$$E[\{(\hat{\omega}/\hat{\alpha}) + \hat{\beta}\} - \{(\omega_0/\alpha_0) + \beta_0)\}]^2.$$

We use simulations to estimate these quantities. Then we considered two data sets, namely, (a) The monthly log returns of IBM stock from 1926 to 1999 (888 observations with first value 1:0434 and last value 4:5633) and (b) The monthly excess returns of S&P 500 from 1926 to 1991 (792 observations with the first value 0.0225 and the last value 0.1116). These data sets were analyzed by Tsay (2010) who fitted various types of conditional heteroscedastic models to them. The data can be found in

http://faculty.chicagobooth.edu/ruey.tsay/teaching/fts3/

We have computed various types of M-estimators for the GARCH and GJR models fitted to these data sets. All computations reported here, except those in Table 3, are carried out using the software R.

3.1. Simulation study

We consider $R = 1000$ replicates each of sample size $n = 500$ from the models (2) and (6) with errors generated from four different distributions, namely, standard normal, scale mixture of normal distributions $(1 - c_*)\Phi(x) + c_*\Phi(x/\sigma)$ with $c_* = 0.05$ and $\sigma^2 = 9$, and standardized student-t distributions with 3 and 4 degrees of freedom. For each sample, we computed five different M-estimators, namely, QMLE, LAD, Huber with $k = 1.5 \times 1.483\,\text{median}|\hat{\epsilon}_t^M|$, B-estimator with $B = 2.5$, and Cauchy; here the pseudo-residuals $\{\hat{\epsilon}_t^M\}$ in the Huber's estimates are defined as in (19) based on MATLAB-prescribed initial estimates $\hat{\alpha}_0^M = 0.05$, $\hat{\beta}_0^M = 0.85$, $\hat{\gamma}_0^M = 0$, and $\hat{\omega}_0^M = (1 - \hat{\alpha}_0^M - \hat{\beta}_0^M) \times \hat{v}(X)$ where $\hat{v}(X)$ is the sample variance of the observed series $\{X_1, \ldots, X_n\}$.

Since the MSEs depend on the underlying true parameter $\boldsymbol{\theta}_0$, we describe a general scenario of relative comparison by reporting a representative result simulated with the values of the true parameters $\omega_0 = 0.005$, $\alpha_0 = 0.2$, and $\beta_0 = 0.75$ for the GARCH (1, 1) model in Table 1. The choice is guided by the estimates of GARCH parameters computed using MATLAB and reported in Table 3 for the S&P 500 data based on QMLE where estimate of ω is very small and those of α and β are moderate with sum close to but less than 1. We report a representative result for a stationary GJR (1, 1) model in Table 2 simulated from $\omega_0 = 0.5$, $\alpha_0 = 0.3$, $\beta_0 = 0.4$, and $\gamma_0 = 0.25$ which are all moderately large and the underlying model is stationary; results corresponding to the other parameter combinations are available from the author upon request.

Tables 1 and 2 show the estimated MSE for each score functions with their standard errors in parentheses computed over R replications; entries in bold represents the least value of MSE for each row. As expected, the QMLE performs well under normal error distribution, but it is not a good choice with other heavy-tailed error densities. Peng and Yao (2003) suggested that when $\{\epsilon_t\}$ follows a heavy-tailed distribution, least absolute deviations estimators (LAD) should be used. Our study reveals that there are score

Table 1
Mean squared error of M-estimators for GARCH (1, 1) model

$n = 500$	QMLE	LAD	Hubers	B-Estimator	Cauchy
MSE			*Normal Distribution*		
	0.0202	0.0677	0.0272	0.0441	0.0812
	(0.0631)	(0.1468)	(0.0758)	(0.1094)	(0.1428)
MSE			*Scale Mixture of Normal Distributions*		
	0.0720	0.0440	0.0434	**0.0408**	0.0909
	(0.1106)	(0.0883)	(0.0879)	(0.0839)	(0.1450)
MSE			*Student-t Distribution (3)*		
	0.0302	0.0163	**0.0119**	0.0133	0.0204
	(0.0745)	(0.0595)	(0.0392)	(0.0488)	(0.0604)
MSE			*Student-t Distribution (4)*		
	0.0241	0.0153	**0.0145**	0.0153	0.0351
	(0.0535)	(0.0448)	(0.0430)	(0.0398)	(0.0925)

Table 2
Mean squared error of M-estimators for GJR (1, 1) model

$n = 500$	QMLE	LAD	Hubers	B-Estimator	Cauchy
MSE			*Normal Distribution*		
	0.0727	0.0304	0.0303	0.0373	0.0371
	(0.0555)	(0.0178)	(0.0187)	(0.0337)	(0.0268)
MSE			*Scale Mixture of Normal Distributions*		
	0.1081	0.0601	0.0593	**0.0454**	0.0566
	(0.0925)	(0.0782)	(0.0811)	(0.0175)	(0.0177)
MSE			*Student-t Distribution (3)*		
	0.0786	0.0400	0.0503	**0.0294**	0.0322
	(0.0625)	(0.0378)	(0.0853)	(0.0154)	(0.0160)
MSE			*Student-t Distribution (4)*		
	0.0815	0.0598	0.0600	**0.0547**	0.0597
	(0.0430)	(0.0315)	(0.0132)	(0.0316)	(0.0727)

functions such as Huber's estimator and B-estimator that can perform even better than the LAD; this comes without imposing any extra restriction such as median$(\epsilon_t^2) = 1$. These results indicate that B-estimator is an excellent choice in terms of the MSE criterion for estimating parameters of GJR and GARCH models when data have heavy tails or there is evidence of outliers. Additional simulations reveal that as the sample size increases, B-estimator performs even better compared to all its competitors in terms of even smaller MSE. Simulation study suggests that $B = 2.5$ works well.

3.2. Financial data

In this section, we compute M-estimates of the parameters by fitting GARCH (1, 1) and GJR (1, 1) models for the centered IBM stock and the centered S & P 500 index, denoted by $\{X_t = r_t - \bar{r}; 1 \le t \le n\}$, where $\{r_t; 1 \le t \le n\}$ is the original stock or index data. Table 4 shows estimated parameters of GARCH (1, 1) model for the IBM data using five different M-estimators and their standard errors (SEs). The Ljung-Box statistics $\{Q(k)\}$ at lag k for the squared residuals $\{\hat{\epsilon_t}^2\}$ are also computed to check the model adequacy. For each estimator, similar and high p-values of the Ljung-Box statistics at lag $k = 10$ suggest that GARCH (1, 1) model is adequate for the data at 5% significance level. As mentioned earlier an M-estimator based on a score function H consistently estimates $\boldsymbol{\theta}_{0H} = (c_H\omega_0, c_H\alpha_0, \beta_0)'$. Note that all M-estimates of β_0 should have similar value free from c_H, and this is reflected in Table 4.

Next, the parameters of the GJR (1, 1) model are estimated for the IBM data, and the M-estimators are reported in Table 5. Standard errors for these estimated parameters are also computed along with the Ljung-Box statistics for $\hat{\epsilon_t}^2$. Based on each estimator, high p-values of the Ljung-Box statistics for lag 10 suggest that GJR (1, 1) model is also adequate for this data set. Moreover, except for the case of Huber's estimator, all other M-estimators do not reject the hypothesis $\gamma = 0$ (that is the GARCH is

Table 3
QMLE of GARCH (1, 1) and GJR (1, 1) parameters for real data sets using MATLAB; SEs are in parantheses

Data Set	IBM Stock		S & P 500 Index	
	GARCH (1,1)	GJR (1,1) $GARCH(1,1)$		$GJR(1,1)$
ω	2.9987	3.3579	0.00008	0.00009
	(0.9415)	(0.9810)	(0.00002)	(0.00002)
α	0.0953	0.0667	0.1211	0.0727
	(0.0201)	(0.0238)	(0.0199)	(0.0210)
γ	–	0.0558	–	0.0822
	–	(0.0256)	–	(0.0283)
β	0.8376	0.8293	0.8556	0.8543
	(0.0365)	(0.0380)	(0.0190)	(0.0185)

Table 4
M-estimates of GARCH (1, 1) parameters and the corresponding Ljung-Box statistic of squared residuals for the IBM data

Parameters	QMLE	LAD	Hubers	B-Estimator	Cauchy
$c_H\omega$	3.0045	1.6319	1.9419	2.0021	0.8984
	(1.4277)	(0.7314)	(0.8795)	(1.0151)	(0.4722)
$c_H\alpha$	0.0950	0.0542	0.0680	0.0717	0.0297
	(0.0307)	(0.0162)	(0.0201)	(0.0236)	(0.0105)
β	0.8378	0.8475	0.8557	0.8502	0.8473
	(0.0535)	(0.0465)	(0.0435)	(0.0502)	(0.0547)
$Q(10)$	2.8528	3.0512	3.2429	3.1591	3.0479
p-value	0.9847	0.9802	0.9751	0.9774	0.9803

adequate) correctly. M-estimators for GJR (1, 1) based on a score function H consistently estimates $\boldsymbol{\theta}_{0H} = (c_H\omega_0, c_H\alpha_0, c_H\gamma_0, \beta_0)'$. Again it is evident from Table 5 that the estimates of β_0 using different score functions are close to each other.

4. Value at risk and M-tests

Next, we consider the prediction of VaR based on M-estimates. A $(1-p)100\%$ VaR is the pth conditional quantile of the distribution of the returns, where p is known and close to zero. Hence, for the returns $\{X_t; 1 \le t \le n\}$ of a portfolio, the VaR at time $t > 1$, denoted by $q_t = q_t(p)$, is defined by

$$q_t = \inf\{x; p \le P_{t-1}(X_t \le x)\},$$

where P_{t-1} is the conditional distribution of X_t given the information available up to time $t-1$. From (15), we get

$$q_t = v_t^{1/2}(\boldsymbol{\theta}_0)F^{-1}(p),$$

Table 5
M-estimates of GJR (1, 1) parameters and the corresponding Ljung-Box statistic of squared residuals for the IBM data

Parameters	QMLE	LAD	Hubers	B-Estimator	Cauchy
$c_H\omega$	3.4542	1.7702	2.2448	2.2262	0.9251
	(1.5490)	(0.7512)	(0.3227)	(1.0468)	(0.4538)
$c_H\alpha$	0.0676	0.0377	0.0471	0.0490	0.0187
	(0.0333)	(0.0173)	(0.0074)	(0.0249)	(0.0105)
$c_H\gamma$	0.0570	0.0373	0.0489	0.0552	0.0255
	(0.0429)	(0.0232)	(0.0100)	(0.0346)	(0.0153)
β	0.8257	0.8383	0.8431	0.8381	0.8412
	(0.0569)	(0.0477)	(0.0156)	(0.0514)	(0.0528)
$Q(10)$	2.8068	3.0582	3.1182	3.2097	3.2548
p-value	0.9856	0.9800	0.9785	0.9761	0.9748

where F^{-1} is the quantile function of the innovations $\{\epsilon_t\}$. From (12) and (14), notice that

$$v_t(\boldsymbol{\theta}_{0H}) = c_H v_t(\boldsymbol{\theta}_0).$$

Hence

$$q_t = c_H^{1/2} v_t^{1/2}(\boldsymbol{\theta}_0) F^{-1}(p)/c_H^{1/2} = v_t^{1/2}(\boldsymbol{\theta}_{0H}) F_*^{-1}(p), \tag{24}$$

where notice that $F_*^{-1}(p)$ is the pth quantile of the scaled errors $\{\epsilon_t/c_H^{1/2}\}$. Estimating $v_t^{1/2}(\boldsymbol{\theta}_{0H})$ by $\widehat{v}_t^{1/2}(\hat{\boldsymbol{\theta}}_n)$ and $F_*^{-1}(p)$ by the pth quantile of the residuals $\{X_t/\{\widehat{v}_t(\hat{\boldsymbol{\theta}}_n)\}^{1/2}; 1 \le t \le T\}$, we obtain from (24) the VaR estimate $\hat{q}_t$ of q_t as

$$\hat{q}_t = \widehat{v}_t^{1/2}(\hat{\boldsymbol{\theta}}_n) \times \big([np]+1\big)\text{th order statistics of } \{X_t/\{\widehat{v}_t(\hat{\boldsymbol{\theta}}_n)\}^{1/2}\}, \quad 2 \le t \le n. \tag{25}$$

Clearly $\hat{q}_t$ depends on the underlying M-estimates.

Let

$$n_* = \sum_{t=2}^{n} I_t \text{ with } I_t = I(X_t \le \hat{q}_t)$$

denote the total number of observed violations. The closeness of the empirical rejection probability

$$\hat{p} = n_*/n \tag{26}$$

to "p" can be used to assess the overall predictive performance of the underlying conditional heteroscedastic model and the M-estimates used for computing $\hat{q}_t$. We describe below two statistical tests for the null hypothesis $E(n_*/n) = p$ against its negation, as they are related to the model validity.

4.1. M-tests

Define the unconditional likelihood ratio test statistic by

$$\mathrm{LR}_{uc} = 2\Big[\ln\big\{(1-\hat{p})^{n-n_*}\hat{p}^{n_*}\big\} - \ln\big\{(1-p)^{n-n_*}p^{n_*}\big\}\Big].$$

Kupiec (1995) proposed this statistics when the QMLE is used as $\hat{\boldsymbol{\theta}}_n$ and the test statistics are asymptotically $\chi^2_{(1)}$.

Note, however, that in a reasonable model of VaR, the previous history of violations should not convey any information about whether or not an additional VaR violations may occur in future. Toward that, using the QMLE as $\hat{\boldsymbol{\theta}}_n$, Christoffersen (1998) defined the independence coverage test statistic, denoted by $\mathrm{LR}_{\mathrm{ind}}$, which characterizes the ways in which these violations occur as follows.

For $i, j = 0, 1$, let n_{ij} be the number of time points $\{t; 2 \le t \le n\}$ for which $I_t = i$ is followed by $I_{t+1} = j$. Let

$$\hat{\pi}_{ij} = n_{ij}/(n_{i0} + n_{i1}), \;\; \hat{\pi} = (n_{01} + n_{11})/n.$$

Then

$$\mathrm{LR}_{ind} = 2\Big[\ln\big((1-\hat{\pi}_{01})^{n_{00}}\hat{\pi}_{01}^{n_{01}}(1-\hat{\pi}_{11})^{n_{10}}\hat{\pi}_{11}^{n_{11}}\big) - \ln\big((1-\hat{\pi})^{(n_{00}+n_{10})}\hat{\pi}^{(n_{01}+n_{11})}\big)\Big].$$

Since both the unconditional coverage and the independence properties should be satisfied for an accurate VaR model, Christoffersen (1998) proposed the statistic

$$\mathrm{LR}_{cc} = \mathrm{LR}_{uc} + \mathrm{LR}_{ind}$$

which is asymptotically $\chi^2_{(2)}$. We consider the same test statistics when $\{\hat{q}_t\}$s are evaluated using M-estimates.

4.1.1. Dynamic quantile M-test

Since the LR_{cc} test only checks the first-order dependence in the risk estimates, Engle and Manganelli (2004) proposed this test to check the high-order dependence among $\{I_t\}$s when the QMLE is used as $\hat{\boldsymbol{\theta}}_n$. To describe it, let the tth response h_t, $2 \le t \le n$, be defined by

$$h_t = \begin{cases} 1-p & \text{if } \; X_t \le \hat{q}_t, \\ -p & \text{if } \; X_t > \hat{q}_t \end{cases}$$

and $h_1 = -p$. Now consider a linear regression model with response $\boldsymbol{Y} = [h_1, \ldots, h_n]'$ and a $n \times k$ design matrix $\boldsymbol{X} = [x_{t,j}]$ with $k = 7$ and all ones in the first column. For the (t, j)th term with $2 \le j \le 6$, $x_{t,j} = h_{t-j}$ if $j < t$ and $x_{t,j} = 0$ if $j \ge t$ and $x_{t,7} = \hat{q}_t$. The dynamic quantile test statistics are defined as

$$\mathrm{DQ} = \frac{\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}}{p(1-p)},$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ is the ordinary least square (OLS) estimator. The DQ test has an asymptotic chi-square distribution with $k = 7$ degrees of freedom under independence.

4.2. Comparisons among competing M-estimators

After assessing model validity using above tests based on different M-estimators, we can make pairwise comparisons of only the competing M-estimators based on VaR in terms of the following two criteria namely the mean relative bias and quadratic loss.

4.2.1. Mean relative bias (MRB)

Suppose there are c number of competing VaR estimates $\{\hat{q}_{jt}; 1 \le t \le n, 1 \le j \le c\}$. Hendricks (1996) defined the mean relative bias (MRB) of the jth estimator ($1 \le j \le c$) as

$$\mathrm{MRB}_j = \frac{1}{n}\sum_{t=1}^{n} \frac{\hat{q}_{jt} - \bar{q}_t}{\bar{q}_t}, \quad \text{where} \quad \bar{q}_t = \frac{1}{c}\sum_{j=1}^{c} \hat{q}_{jt}.$$

4.2.2. Average quadratic loss (AQL)

The statistic n_* based on a particular estimator or method counts merely the number of violations and does not consider the magnitude of losses. To take into account this, Lopez (1999) defined the overall quadratic loss of a VaR estimate by $\sum_{t=1}^{n} L_t/n$ where

$$L_t = \begin{cases} 1 + (\hat{q}_t - X_t)^2 & \text{if } X_t \le \hat{q}_t, \\ 0 & \text{if } X_t > \hat{q}_t. \end{cases}$$

We can use the loss corresponding to different estimates to compare their performance.

5. Data analysis based on VaR

In this section, we exhibit the robustness and better performance of the above M-estimators by demonstrating that irrespective of the form GARCH or more general GJR assumed, the VaR estimates based on different M-estimators are accurate subject to sampling variations. Moreover, minimum AQL is incurred when Huber, B-estimator, or the Cauchy estimator is used. The data sets used in the empirical application are the daily closing indices $\{P_t\}$ of three major stocks of the United States, Europe, and Asia, namely, S&P500 Index, FTSE100 Index, and NIKKEI225 Index, respectively. The data sets are obtained for the period of January 1990 to December 2005 from the website *http://www.finance.yahoo.com*. Note that S&P500 and FTSE100 indices consist of total $T = 4042$ values whereas the NIKKEI225 Index consists of $T = 3938$ observations. For each of the three indices, the return at time t is defined as

$$r_t = (\ln P_t - \ln P_{t-1}) \times 100\%, \quad t = 1, 2, \ldots, n.$$

Next using $\{X_t = r_t - \bar{r}; 1 \le t \le n\}$ (with $\bar{r} = \sum_{t=1}^{n} r_t/n$) as our observations, each data series is divided into two parts; the estimation or in-sample part of initial K values and the validation or out-of-sample part of $N = n - K$ values. For this study, we fix $N = 2000$ for each data set.

We fitted both GJR and GARCH models to all three data sets; note that GARCH is a special case of GJR model. For evaluating the accuracy of VaR estimates, we present

below the results of some continual statistical testing (backtesting) on the estimation sample. Backtesting helps to identify the validity of each model and is also required by the regulatory bodies such as the Basel Committee on Banking Supervision (1996).

5.1. *In-sample VaR evaluation and comparison*

In this subsection, we assume that the sample size is merely K. Note that the K-values for the S&P500, FTSE100, and NIKKEI225 are 2042, 2042, and 1938, respectively. We compute $K - 1$ numbers of in-sample VaR estimates with $p = 10\%$ using (25). Subsequently, we compute all statistics of Section 6.

Tables 6 and 7 report the results of in-sample VaR estimates. For brevity, we report results corresponding to $p = 10\%$ only though more simulation results corresponding to other values of p are available in the study by Iqbal (2010). The first row for each data set shows that the $\{\hat{p}\}$'s are quite close to p for all M-estimators for both models indicating that both GJR and its special case GARCH fit these real data sets well; thus, the VaR estimates are robust to the model specification for symmetry. Next we explore performance of various M-test statistics to check model validity and at the same time we analyze their ability to detect model misspecification. None of the coverage statistics LR_{uc} and LR_{cc} is statistically significant, which indicates that the expected and the actual proportion of observations falling below the VaR threshold remain statistically same for both models. However, for the FTSE100 and NIKKEI225 returns, although all M-estimators pass the conditional coverage statistics at both $p = 5\%$ and 10%, they fail to accept the null based on the dynamic quantile test at these rejection probabilities when GARCH model is used alone showing the existence of high-order dependence. However, with the use of more general GJR, it is no longer significant pointing toward the need of using asymmetric model.

After noting that all estimators have passed the coverage tests for both models, we present comparisons of the competing estimators. The AQL for each estimator is reported in Tables 8 and 9. It turns out that AQL is a robust measure with respect to the choice of models. For both models and for all three data sets, the least AQLs are related to the use of Huber, Cauchy, and B-estimator. For example, B-estimator produced the least AQL for NIKKEI225 Index for both models. Moreover, the signs of MRB are consistent with both models for all three data sets. Thus, our analysis reveals the existence of alternative estimators that perform better than the QMLE in the VaR evaluation and whose performance is robust with respect to the choice of symmetric or asymmetric models.

5.2. *Out-of-sample VaR evaluation and comparison*

Next, we look at the performance of M-estimators in producing one-step-ahead VaR estimates; here we report results corresponding to $p = 10\%$. In this setup, we allow the set of observations of size K to change over time using moving window of length K and producing altogether N number of VaR estimates. This is what is implemented in the "real-life" situation where the out-of-sample VaR forecasts are delivered based on

Table 6
In-sample M-statistics for the VaR evaluation and model validity (GARCH)

	QMLE	LAD	Huber	B-Estimator	Cauchy
			90% VaR Confidence Level		
S&P500 Index					
$\hat{p}$	0.0950	0.0955	0.0955	0.0970	0.0945
LR_{uc}	0.5747	0.4668	0.4668	0.2111	0.6940
LR_{cc}	5.1378	4.8136	4.8136	4.8616	4.5303
DQ	10.4852	9.9082	11.2013	10.9010	10.2724
FTSE100 Index					
$\hat{p}$	0.0955	0.0955	0.0940	0.0945	0.0926
LR_{uc}	0.4668	0.4668	0.8247	0.6940	1.2860
LR_{cc}	1.1401	0.6940	1.0225	0.8968	1.4974
DQ	16.1487*	20.4942**	23.0526**	26.9402**	28.3048**
NIKKEI225 Index					
$\hat{p}$	0.1042	0.1022	0.1022	0.1017	0.1017
LR_{uc}	0.3808	0.1005	0.1005	0.0584	0.0584
LR_{cc}	0.5987	3.7972	3.7972	3.9585	3.9585
DQ	19.0729**	23.4811**	23.3008**	24.3819**	24.4052**

Note: The DQ test statistic is asymptotically $\chi^2(7)$, and * and ** denote significance at the 5% and 1% level, respectively.

Table 7
In-sample M-statistics for the VaR evaluation and model validity (GJR)

	QMLE	LAD	Huber	B-Estimator	Cauchy
			90% VaR Confidence Level		
S&P500 Index					
$\hat{p}$	0.0965	0.0975	0.0955	0.0989	0.0960
LR_{uc}	0.2851	0.1483	0.4668	0.0264	0.3703
LR_{cc}	1.9906	3.6884	2.4323	2.3053	2.2031
DQ	8.6549	10.6801	11.4062	12.3916	10.1888
FTSE100 Index					
$\hat{p}$	0.0960	0.0960	0.0950	0.0955	0.0960
LR_{uc}	0.3703	0.3703	0.5747	0.4668	0.3703
LR_{cc}	1.1124	0.7939	0.7873	3.8900	4.4786
DQ	9.6957	9.1298	11.8438	10.6586	13.7654
NIKKEI225 Index					
$\hat{p}$	0.1037	0.1037	0.1032	0.1027	0.1042
LR_{uc}	0.2940	0.2940	0.2183	0.1538	0.3808
LR_{cc}	2.6928	1.5418	2.1247	1.1521	1.1304
DQ	9.8554	8.4766	9.5726	6.4889	6.7659

Note: The DQ test statistic is asymptotically $\chi^2(7)$, and * and ** denote significance at the 5% and 1% level, respectively.

Table 8
Comparison among competing M-estimators for the in-sample VaR evaluation (GARCH)

	QMLE	LAD	Huber	B-Estimator	Cauchy
			90% VaR Confidence Level		
S&P500 Index					
MRB	0.0063	0.0046	–0.0033	–0.0082	0.0006
AQL	0.1069	0.1073	0.1074	0.1089	**0.1062**
FTSE100 Index					
MRB	0.0140	–0.0049	–0.0026	–0.0064	–0.0001
AQL	0.1019	0.1021	0.1007	0.1012	**0.0992**
NIKKEI225 Index					
MRB	–0.0007	0.0014	–0.0031	0.0023	0.0001
AQL	0.1311	0.1286	0.1291	**0.1280**	0.1282

Note: The smallest AQL for each data set is bold faced to highlight the best performance.

Table 9
Comparison among competing M-estimators for the in-sample VaR evaluation (GJR)

	QMLE	LAD	Huber	B-Estimator	Cauchy
			90% VaR Confidence Level		
S&P500 Index					
MRB	0.0036	0.0079	–0.0035	–0.0090	0.0011
AQL	0.1083	0.1089	**0.1072**	0.1108	0.1076
FTSE100 Index					
MRB	0.0010	–0.0026	–0.0051	–0.0019	–0.0096
AQL	0.1024	0.1025	**0.1016**	0.1023	0.1031
NIKKEI225 Index					
MRB	0.0109	0.0001	–0.0003	–0.0025	–0.0082
AQL	0.1272	0.1275	0.1266	**0.1259**	0.1279

Note: The smallest AQL for each data set is bold faced to highlight the best performance.

the estimated parameters of the daily updated observations. In other words, for each of $1 \leq w \leq N$, we produce one-step-ahead VaR estimate $\hat{q}_{ow}$ based on $\{X_t; w \leq t \leq w + K - 1\}$ using (25).

Tables 10 and 11 provide results related to the out-of-sample VaR estimates. The empirical rejection probability of (26) for the out-of-sample VaR estimates is defined as

$$\hat{p}_o = (1/N) \sum_{w=1}^{N} I(X_{w+K} \leq \hat{q}_{ow}).$$

For all data sets and M-estimators, $\hat{p}_o$ is close to $p = 0.10$ for both models indicating robustness of the VaR estimates with respect to the models used. Neither of the likelihood ratio statistics, LR_{uc} and LR_{cc}, is significant at 10% and 5% levels showing that the proportion of violations produced by M-estimators is statistically same as the

Table 10
Out-of-sample M-statistics for the VaR evaluation and model validity (GARCH)

	QMLE	LAD	Huber	B-Estimator	Cauchy
			90% VaR Confidence Level		
S&P500 Index					
$\hat{p}_o$	0.1120	0.1120	0.1110	0.1140	0.1090
LR_{uc}	3.0927	3.0927	2.6058	4.1867	1.7541
LR_{cc}	3.7433	4.0705	4.2349	4.6202	3.3588
DQ	16.2397*	21.7491**	22.6867**	18.4834**	14.7659*
FTSE100 Index					
$\hat{p}_o$	0.1055	0.1050	0.1055	0.1045	0.1060
LR_{uc}	0.6616	0.5475	0.6616	0.4441	0.7862
LR_{cc}	4.7973	4.0543	3.9799	4.1451	3.9221
DQ	19.1367**	19.1286**	20.1013**	19.8089**	18.5360**
NIKKEI225 Index					
$\hat{p}_o$	0.1005	0.1005	0.0995	0.0995	0.0990
LR_{uc}	0.0055	0.0055	0.0056	0.0056	0.0223
LR_{cc}	0.2202	0.2552	0.2175	0.2566	0.3497
DQ	7.3476	5.3039	5.6436	5.4985	4.2406

Note: The DQ test statistic is asymptotically $\chi^2(7)$, and * and ** denote significance at the 5% and 1% level, respectively.

Table 11
Out-of-sample M-statistics for the VaR evaluation and model validity (GJR)

	QMLE	LAD	Huber	B-Estimator	Cauchy
			90% VaR Confidence Level		
S&P500 Index					
$\hat{p}_o$	0.1100	0.1105	0.1110	0.1110	0.1130
LR_{uc}	2.1595	2.3776	2.6058	2.6058	3.6197
LR_{cc}	2.5563	2.7369	3.1171	3.1171	3.9627
DQ	9.3568	11.9913	13.8734	13.3806	14.9990*
FTSE100 Index					
$\hat{p}_o$	0.1060	0.1090	0.1070	0.1055	0.1070
LR_{uc}	0.7862	1.7541	1.0671	0.6616	1.0671
LR_{cc}	3.9221	2.8918	3.8556	3.2526	3.8556
DQ	19.7151**	11.0671	14.9138*	14.2534*	11.9589
NIKKEI225 Index					
$\hat{p}_o$	0.0995	0.1005	0.0980	0.0985	0.0995
LR_{uc}	0.0056	0.0055	0.0894	0.0502	0.0056
LR_{cc}	0.7268	0.8766	0.9901	1.0344	0.4027
DQ	7.4248	6.4519	5.6769	5.4136	4.8600

Note: The DQ test statistic is asymptotically $\chi^2(7)$, and * and ** denote significance at the 5% and 1% level, respectively.

expected proportion p. However, at $p = 10\%$, the DQ statistics for all estimators fail to accept the null hypothesis of no higher order dependence in VaR violations in S&P500 when the GARCH model is fitted. This disappears when more general GJR is fitted. From these tables, similar to the in-sample VaR, for the the FTSE100 and NIKKEI225

indices, the AQL is least in connection with Huber, Cauchy, and B-estimator for both models. However, for S&P500, the AQL is unexpectedly least when QMLE is used for GJR model.

6. Nonlinear AR–ARCH model

Autoregressive models with heteroscedastic errors appear quite often in practice. Consider, for example, observations $\{X_i; 0 \le i \le n\}$ satisfying

$$X_i = \alpha X_{i-1} + \{\beta_0 + \beta_1 X_{i-1}^2\}^{1/2} \eta_i, \quad 1 \le i \le n, \tag{27}$$

where $\alpha \in I\!R$, $\boldsymbol{\beta} = (\beta_0, \beta_1)' \in (0, \infty)^2$, and $\{\eta_i\}$s are i.i.d. with zero mean and unit variance. This is called an AR(1)–ARCH(1) model. Another interesting model is

$$X_i = \alpha X_{i-1} + \{\beta_1 X_{i-1} I(X_{i-1} > 0) - \beta_2 X_{i-1} I(X_{i-1} \le 0)\} \eta_i, \quad 1 \le i \le n, \tag{28}$$

where $\alpha \in I\!R$, $\boldsymbol{\beta} = (\beta_0, \beta_1)' \in (0, \infty)^2$, and $\{\eta_i\}$s are i.i.d. with zero mean and unit variance. This can be used to model asymmetric feature of volatility where the effect of a positive news is β_1, whereas that of a negative news is β_2. In this section, we are primarily interested in the estimation of the mean parameter α as opposed to the previous sections where we considered the estimation of the variance parameters.

To motivate the estimator of α in (27), first note that

$$E[\{\beta_0 + \beta_1 X_{i-1}^2\}^{1/2} \eta_i] = E[\{\beta_0 + \beta_1 X_{i-1}^2\}^{1/2}] E[\eta_i] = 0,$$

and hence, ignoring the heteroscedasticity, one can estimate the mean parameter α using simple least squares by

$$\hat{\alpha}_p = \left[\sum_{i=1}^{n} X_{i-1}^2\right]^{-1} \sum_{i=1}^{n} X_i X_{i-1}.$$

This estimator is consistent but clearly inefficient. However, using M-estimators discussed in previous sections with observations $\{X_i - \hat{\alpha}_p X_{i-1}; 1 \le i \le n\}$, one can obtain estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ of the heteroscedastic parameters. Using these estimators, an approximation of the model (27) is

$$\frac{X_i}{\{\hat{\beta}_0 + \hat{\beta}_1 X_{i-1}^2\}^{1/2}} \approx \alpha \frac{X_{i-1}}{\{\hat{\beta}_0 + \hat{\beta}_1 X_{i-1}^2\}^{1/2}} + \eta_i,$$

which has homoscedastic errors. Now using standard robust estimation methods for homoscedastic linear regression and autoregression, one can obtain improved estimators of α.

To explain this procedure in the general context, note that both models (27) and (28) can be cast into the following general framework of the nonlinear autoregressive

model with autoregressive conditional heteroscedastic errors. Suppose that s, p, r_1, and r_2 are known integers and $\{X_i,\ 1-s \le i \le n\}$ is an observable time series. For $1 \le i \le n$, set $\boldsymbol{W}_{i-1} \coloneqq (X_{i-1}, X_{i-2}, \ldots, X_{i-s})'$, the vector of lagged observations. To achieve bit more generality, let $\boldsymbol{Y}_{i-1} = c(\boldsymbol{W}_{i-1})$, where c: $\mathbb{R}^s \to \mathbb{R}^p$ is a known function; for most applications, $s = p$ and c is the identity function. Let Ω_j, $j = 1, 2$, be open subsets of $\mathbb{R}^{r_1}$ and $\mathbb{R}^{r_2}$, respectively; they are the parameter spaces. Let μ and σ be known functions, respectively, from $\mathbb{R}^p \times \Omega_1$ to $\mathbb{R}$ and $\mathbb{R}^p \times \Omega_2$ to $\mathbb{R}^+ \coloneqq (0, \infty)$, which are differentiable in their second argument. Consider a model where for some $\boldsymbol{\alpha} \in \Omega_1$, $\boldsymbol{\beta} \in \Omega_2$,

$$\eta_i = \{X_i - \mu(\boldsymbol{Y}_{i-1}, \boldsymbol{\alpha})\}/\sigma(\boldsymbol{Y}_{i-1}, \boldsymbol{\beta}) \quad 1 \le i \le n$$

are i.i.d. with mean zero, variance 1 and independent of $\boldsymbol{W}_0 \coloneqq (X_0, X_{-1}, \cdots, X_{1-s})'$. In other words, the observations satisfy

$$X_i = \mu(\boldsymbol{Y}_{i-1}, \boldsymbol{\alpha}) + \sigma(\boldsymbol{Y}_{i-1}, \boldsymbol{\beta})\, \eta_i, \quad i \ge 1, \tag{29}$$

where the errors $\{\eta_i,\ i \ge 1\}$ are independent of Y_0, and i.i.d. standard random variables having a distribution function G and the density function g.

In the following, we cite some examples of (29).

Example 8 (Engle's ARCH Model). In the ARCH model introduced by Engle (1982), one observes $\{Z_i, 1-s \le i \le n\}$ such that

$$Z_i = (\alpha_0 + \alpha_1 Z_{i-1}^2 + \ldots + \alpha_s Z_{i-s}^2)^{1/2} \varepsilon_i, \quad 1 \le i \le n, \tag{30}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_s)' \in \mathbb{R}^{+(s+1)} \coloneqq (0, \infty)^{(s+1)}$ is the unknown parameter and $\{\varepsilon_i; 1 \le i \le n\}$ are unobservable i.i.d. with mean zero, variance 1 and finite fourth moment.

Squaring both sides of (30) and writing $\eta_i \coloneqq \varepsilon_i^2 - 1$, $X_i = Z_i^2$, $\boldsymbol{W}_{i-1} = [X_{i-1}, \ldots, X_{i-s}]' = [Z_{i-1}^2, \ldots, Z_{i-s}^2]'$, and $\boldsymbol{Y}_{i-1}' = [1, \boldsymbol{W}_{i-1}']$, model (30) can be recast as

$$X_i = \boldsymbol{Y}_{i-1}'\boldsymbol{\alpha} + (\boldsymbol{Y}_{i-1}'\boldsymbol{\alpha})\, \eta_i, \quad 1 \le i \le n. \tag{31}$$

(31) is an example of the model (29) with $\boldsymbol{\alpha} = \boldsymbol{\beta}$, $c(\boldsymbol{w}) = [1, \boldsymbol{w}]'$, $\boldsymbol{w} \in [0, \infty)^s$, $p = s + 1$, $r_1 = r_2 = s + 1$, and

$$\mu(\boldsymbol{y}, \boldsymbol{a}) = \boldsymbol{y}'\boldsymbol{a},\ \sigma(\boldsymbol{y}, \boldsymbol{b}) = \boldsymbol{y}'\boldsymbol{b}.$$

Example 9 (Autoregressive Linear Square Conditional Heteroscedastic Model – ARLSCH). Consider the first-order autoregressive model with heteroscedastic errors where one observes $\{X_i; 0 \le i \le n\}$ such that the conditional variance of the ith observation X_i depends linearly on the squares of past as follows:

$$X_i = \alpha X_{i-1} + \{\beta_0 + \beta_1 X_{i-1}^2\}^{1/2} \eta_i, \quad 1 \le i \le n, \tag{32}$$

where $\alpha \in I\!R$, $\boldsymbol{\beta} = (\beta_0, \beta_1)' \in (0,\infty)^2$, and $\{\eta_i\}$s are i.i.d. with zero mean and unit variance. With the identification $s = 1 = p$, $c(w) = w$, $r_1 = 1$, $r_2 = 2$, and

$$\mu(y,a) = ya,\ \sigma(y,\boldsymbol{b}) = (b_0 + b_1 y^2)^{1/2},\ y \in I\!R,$$

model (32) can be seen as an example of (29).

The assumption needed on the parameters under which the process $\{X_i; i \geq 0\}$ of (32) is strictly stationary and ergodic is as follows:

$$|\alpha| + E|\eta_1| \max\{\beta_0^{1/2}, \beta_1^{1/2}\} < 1.$$

This follows by using Lemma 1 of Härdle and Tsybakov (1997, p. 227) with $C_1 = |\alpha|$ and $C_2 = \max\{\beta_0^{1/2}, \beta_1^{1/2}\} = \sup\{(\beta_0 + \beta_1 x^2)^{1/2}/(1 + |x|); x \in I\!R\}$.

Example 10 (Autoregressive Threshold Conditional Heteroscedastic Model – ARTCH). Consider an sth order autoregressive model with self-exciting threshold heteroscedastic errors, where the conditional standard deviation of the ith observation X_i is piecewise linear on the past as follows:

$$X_i = (\alpha_1 X_{i-1} + \ldots \alpha_s X_{i-s}) + \Big\{\beta_1 X_{i-1} I(X_{i-1} > 0) - \beta_2 X_{i-1} I(X_{i-1} \leq 0) \cdots$$
$$+ \beta_{2s-1} X_{i-s} I(X_{i-s} > 0) - \beta_{2s} X_{i-s} I(X_{i-s} \leq 0)\Big\}\ \eta_i, \quad 1 \leq i \leq n, \qquad (33)$$

where all β_js are positive and $\{\eta_i\}$s are i.i.d. with zero mean and unit variance. For applications and many probabilistic properties of this model including conditions for the stationarity and ergodicity of $\{X_i\}$, see the work done by Rabemananjara and Zakoian (1993). For a discussion on the difficulties associated with the asymptotics of the robust estimation in this model, due to the lack of differentiability caused by threshold, see the study by Rabemananjara and Zakoian, 1993, p. 38.

With the identification $p = s$, $c(\boldsymbol{w}) = \boldsymbol{w}$, $r_1 = s$, $r_2 = 2s$, and

$$\mu(\boldsymbol{y},\boldsymbol{a}) = \boldsymbol{y}'\boldsymbol{a},\ \sigma(\boldsymbol{y},\boldsymbol{b}) = \sum_{j=1}^{s} b_{2j-1} y_j I(y_j > 0) + \sum_{j=1}^{s} b_{2j}(-y_j) I(y_j \leq 0),$$
$$\boldsymbol{y} \in I\!R^s,\ \boldsymbol{t} \in (0,\infty)^{2s},$$

the model (33) can be seen as an example of (29).

6.1. *M- and R-estimators*

Let $\boldsymbol{\tau} = (\boldsymbol{a},\boldsymbol{b})$ denote a generic value in the parameter space $\Omega_1 \times \Omega_2$ and let $\boldsymbol{\theta} = (\boldsymbol{\alpha},\boldsymbol{\beta})$ be the true parameter. To estimate $\boldsymbol{\alpha}$, we proceed in three steps. Using $E\{\sigma(Y_{i-1},\boldsymbol{\beta})\,\eta_i\} = 0$ in (29), we first propose a preliminary estimator $\widehat{\boldsymbol{\alpha}}_p$; note that the proposal does not take into account the heteroscedasticity of the model, and hence, it gives a consistent but inefficient estimator. Next, we use $\widehat{\boldsymbol{\alpha}}_p$ to construct an estimator $\widehat{\boldsymbol{\beta}}$ of the parameter $\boldsymbol{\beta}$. Finally, substituting $\widehat{\boldsymbol{\alpha}}_p$ and $\widehat{\boldsymbol{\beta}}$ in (29), the heteroscedastic

model is transformed to an approximate nonlinear homoscedastic autoregressive model (36), and we use standard robust estimation procedures of the homoscedastic models to propose improved estimator of $\boldsymbol{\alpha}$.

In the sequel, $\dot{\mu}$ and $\dot{\sigma}$ denote the derivatives of the functions μ and σ, respectively, with respect to their second arguments. Also for a vector $\boldsymbol{y}$, its jth coordinator is denoted as $\boldsymbol{y}_j$.

Step 1: Define

$$\mathcal{H}(\boldsymbol{a}) \coloneqq n^{-1/2} \sum_{i=1}^{n} \dot{\mu}(\boldsymbol{Y}_{i-1}, \boldsymbol{a})\{X_i - \mu(\boldsymbol{Y}_{i-1}, \boldsymbol{a})\}.$$

Since $E[\mathcal{H}(\boldsymbol{\alpha})] = 0$, we define a preliminary estimator $\widehat{\boldsymbol{\alpha}}_p$ of $\boldsymbol{\alpha}$ by the relation

$$\widehat{\boldsymbol{\alpha}}_p \coloneqq \operatorname{argmin} \left\{ \sum_{j=1}^{r_1} |\mathcal{H}_j(\boldsymbol{a})|; \boldsymbol{a} \in \Omega_1 \right\}, \tag{34}$$

where $\mathcal{H}_j(\boldsymbol{a})$ is the jth coordinate of the vector $\mathcal{H}(\boldsymbol{a})$, $1 \le j \le r_1$.

In particular, when $\mu(\boldsymbol{y}, \boldsymbol{a}) = \boldsymbol{y}'\boldsymbol{a}$,

$$\widehat{\boldsymbol{\alpha}}_p = \left[\sum_{i=1}^{n} \boldsymbol{Y}_{i-1} \boldsymbol{Y}'_{i-1} \right]^{-1} \left[\sum_{i=1}^{n} X_i \boldsymbol{Y}_{i-1} \right].$$

Step 2: Let

$$\eta_i(\boldsymbol{\tau}) \coloneqq \{X_i - \mu(\boldsymbol{Y}_{i-1}, \boldsymbol{a})\}/\sigma(\boldsymbol{Y}_{i-1}, \boldsymbol{b}), \ 1 \le i \le n,$$

denote the ith residual. Let κ be a nondecreasing right continuous function on $\mathbb{R}$ such that $E\{\eta_1 \kappa(\eta_1)\} = 1$. This is automatically satisfied, for example, when κ is the identity function ($\kappa(x) \equiv x$). Consider the statistic

$$M_s(\boldsymbol{\tau}) \coloneqq n^{-1/2} \sum_{i=1}^{n} \frac{\dot{\sigma}(\boldsymbol{Y}_{i-1}, \boldsymbol{b})}{\sigma(\boldsymbol{Y}_{i-1}, \boldsymbol{b})} \Big[\eta_i(\boldsymbol{\tau}) \kappa(\eta_i(\boldsymbol{\tau})) - 1 \Big].$$

Since $E[M_s(\boldsymbol{\alpha}, \boldsymbol{\beta})] = 0$, an estimator of the scale parameter $\boldsymbol{\beta}$ is defined by the relation

$$\widehat{\boldsymbol{\beta}} \coloneqq \operatorname{argmin} \left\{ \sum_{j=1}^{r_2} |M_{sj}(\widehat{\boldsymbol{\alpha}}_p, \boldsymbol{b})|; \boldsymbol{b} \in \Omega_2 \right\}.$$

Note that (29) can be written as

$$X_i/\sigma(\boldsymbol{Y}_{i-1}, \boldsymbol{\beta}) = \mu(\boldsymbol{Y}_{i-1}, \boldsymbol{\alpha})/\sigma(\boldsymbol{Y}_{i-1}, \boldsymbol{\beta}) + \eta_i. \tag{35}$$

This in turn can be approximated by

$$X_i/\sigma(\boldsymbol{Y}_{i-1},\widehat{\boldsymbol{\beta}}) \approx \mu(\boldsymbol{Y}_{i-1},\boldsymbol{\alpha})/\sigma(\boldsymbol{Y}_{i-1},\widehat{\boldsymbol{\beta}}) + \eta_i, \tag{36}$$

which is a nonlinear autoregressive model with homoscedastic errors.

Now using the standard definition for homoscedastic nonlinear model (35), the class of M-estimators and R-estimators based on appropriate score functions ψ and φ, respectively, can be defined as follows; see the study by Bose and Mukherjee (2003) for a similar two-step idea.

Step 3: Let ψ be nondecreasing and bounded function on $\mathbb{R}$ such that $E\{\psi(\eta_1)\} = 0$. An example is the function $\psi(x) = sign(x)$ when $\{\eta_i\}$s are symmetrically distributed around 0.

Let $\varphi : [0,1] \to \mathbb{R}$ belong to the class

$$\mathcal{F} = \{\varphi; \varphi\colon [0,1] \to \mathbb{R} \text{ is right continuous, nondecreasing, with } \varphi(1) - \varphi(0) = 1\}.$$

An example of the function belonging to this class is $\varphi(u) = u - 1/2$; it is called the Wilcoxon rank score function. Define the M-statistics

$$M_\psi(\boldsymbol{\tau}) = n^{-1/2} \sum_{i=1}^{n} \frac{\dot{\mu}(\boldsymbol{Y}_{i-1},\boldsymbol{a})}{\sigma(\boldsymbol{Y}_{i-1},\boldsymbol{b})} \psi\{\eta_i(\boldsymbol{\tau})\}.$$

Since $E[M_\psi(\boldsymbol{\alpha},\boldsymbol{\beta})] = 0$, from (35), an M estimator of $\boldsymbol{\alpha}$ corresponding to the score function ψ is defined as

$$\widehat{\boldsymbol{\alpha}}_M := \text{argmin} \left\{ \sum_{j=1}^{r_1} |M_{\psi j}(\boldsymbol{a},\widehat{\boldsymbol{\beta}})|; \boldsymbol{a} \in \Omega_1 \right\}.$$

Define the rank statistic as

$$\boldsymbol{S}_\varphi(\boldsymbol{\tau}) = n^{-1/2} \sum_{i=1}^{n} \left[\frac{\dot{\mu}(\boldsymbol{Y}_{i-1},\boldsymbol{a})}{\sigma(\boldsymbol{Y}_{i-1},\boldsymbol{b})} - n^{-1} \times \sum_{j=1}^{n} \left\{ \frac{\dot{\mu}(\boldsymbol{Y}_{j-1},\boldsymbol{a})}{\sigma(\boldsymbol{Y}_{j-1},\boldsymbol{b})} \right\} \right] \varphi\left(\frac{R_i\boldsymbol{\tau}}{n+1} \right), \boldsymbol{\tau} \in \Omega,$$

where $R_i\boldsymbol{\tau} = \sum_{j=1}^{n} I\{\eta_j(\boldsymbol{\tau}) \le \eta_i(\boldsymbol{\tau})\}$, the rank of $\eta_i(\boldsymbol{\tau})$ among $\{\eta_j(\boldsymbol{\tau}); 1 \le j \le n\}$. Hence, $E[\boldsymbol{S}_\varphi(\boldsymbol{\alpha},\boldsymbol{\beta})] = 0$ and so a generalized R-estimator of $\boldsymbol{\alpha}$ corresponding to the score function φ is defined as

$$\widehat{\boldsymbol{\alpha}}_R = \text{argmin}\{ \left\{ \sum_{j=1}^{r_1} |S_{\varphi j}(\boldsymbol{a},\widehat{\boldsymbol{\beta}})|; \boldsymbol{a} \in \Omega_1 \right\}.$$

6.2. Asymptotic distribution

Define the normalized derivatives

$$\dot{\mu}_i = \frac{\dot{\mu}(\boldsymbol{Y}_{i-1}, \boldsymbol{\alpha})}{\sigma(\boldsymbol{Y}_{i-1}, \boldsymbol{\beta})} \text{ and } \dot{\sigma}_i = \frac{\dot{\sigma}(\boldsymbol{Y}_{i-1}, \boldsymbol{\alpha})}{\sigma(\boldsymbol{Y}_{i-1}, \boldsymbol{\beta})}.$$

We assume the existence of positive definite matrices $\Lambda(\boldsymbol{\theta})$, $\Lambda_c(\boldsymbol{\theta})$ $\boldsymbol{G}(\boldsymbol{\theta})$, and $\boldsymbol{G}_c(\boldsymbol{\theta})$ such that

$$n^{-1}\sum_{i=1}^{n} \dot{\mu}_i \dot{\mu}_i' = \Lambda(\boldsymbol{\theta}) + o_p(1), \quad n^{-1}\sum_{i=1}^{n} \dot{\mu}_i \dot{\sigma}_i' = \boldsymbol{G}(\boldsymbol{\theta}) + o_p(1), \tag{37}$$

and

$$n^{-1}\sum_{i=1}^{n} \left(\dot{\mu}_i - n^{-1}\sum_{i=1}^{n} \dot{\mu}_i\right) \dot{\mu}_i' = \Lambda_c(\boldsymbol{\theta}) + o_p(1),$$

$$n^{-1}\sum_{i=1}^{n} \left(\dot{\mu}_i - n^{-1}\sum_{i=1}^{n} \dot{\mu}_i\right) \dot{\sigma}_i' = \boldsymbol{G}_c(\boldsymbol{\theta}) + o_p(1). \tag{38}$$

When $\{X_i\}$ is stationary and ergodic, such matrices exist under the finiteness of moments of appropriate order with the following expressions

$$\Lambda(\boldsymbol{\theta}) = E\left[\left\{\frac{\dot{\mu}(\boldsymbol{Y}_0, \boldsymbol{\alpha})}{\sigma(\boldsymbol{Y}_0, \boldsymbol{\beta})}\right\}\left\{\frac{\dot{\mu}(\boldsymbol{Y}_0, \boldsymbol{\alpha})}{\sigma(\boldsymbol{Y}_0, \boldsymbol{\beta})}\right\}'\right], \quad \boldsymbol{G}(\boldsymbol{\theta}) = E\left[\left\{\frac{\dot{\mu}(\boldsymbol{Y}_0, \boldsymbol{\alpha})}{\sigma(\boldsymbol{Y}_0, \boldsymbol{\beta})}\right\}\left\{\frac{\dot{\sigma}(\boldsymbol{Y}_0, \boldsymbol{\alpha})}{\sigma(\boldsymbol{Y}_0, \boldsymbol{\beta})}\right\}'\right],$$

$$\Lambda_c(\boldsymbol{\theta}) = \Lambda(\boldsymbol{\theta}) - E\left\{\frac{\dot{\mu}(\boldsymbol{Y}_0, \boldsymbol{\alpha})}{\sigma(\boldsymbol{Y}_0, \boldsymbol{\beta})}\right\} E\left\{\frac{\dot{\mu}(\boldsymbol{Y}_0, \boldsymbol{\alpha})}{\sigma(\boldsymbol{Y}_0, \boldsymbol{\beta})}\right\}',$$

and

$$\boldsymbol{G}_c(\boldsymbol{\theta}) = \boldsymbol{G}(\boldsymbol{\theta}) - E\left\{\frac{\dot{\mu}(\boldsymbol{Y}_0, \boldsymbol{\alpha})}{\sigma(\boldsymbol{Y}_0, \boldsymbol{\beta})}\right\} E\left\{\frac{\dot{\sigma}(\boldsymbol{Y}_0, \boldsymbol{\alpha})}{\sigma(\boldsymbol{Y}_0, \boldsymbol{\beta})}\right\}'. \tag{39}$$

THEOREM 2.

(i) If (37) holds, $\int |x| g(x) d\psi < \infty$ and $\int g d\psi > 0$, then

$$\int g d\psi \, n^{1/2}(\widehat{\boldsymbol{\alpha}}_M - \boldsymbol{\alpha}) = -\Lambda^{-1}(\boldsymbol{\theta}) \Bigg[M_\psi(\boldsymbol{\theta}) + \boldsymbol{G}(\boldsymbol{\theta}) n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \times \int x g(x) d\psi(x) \Bigg] + o_p(1).$$

(ii) Moreover, if either $\int x g(x) d\psi(x) = 0$ or $\boldsymbol{G}(\boldsymbol{\theta}) = 0$, then

$$n^{1/2}(\widehat{\boldsymbol{\alpha}}_M - \boldsymbol{\alpha}) \Rightarrow \mathcal{N}_{r_1}\Big[\boldsymbol{0}, \Lambda^{-1}(\boldsymbol{\theta}) J_M(\psi, G)\Big], \tag{40}$$

where $J_M(\psi, G) = \frac{\int \psi^2(x) g(x) dx}{(\int g d\psi)^2}$.

A sufficient condition for $\int xg(x)d\psi(x) = 0$ is that g is an even and ψ is an odd function.

THEOREM 3.

(i) If (38) holds, $\int |x|g(x)\varphi(G(dx)) < \infty$ *and* $\int g(x)\varphi(G(dx)) > 0$, *then*

$$\int g(x)\varphi(G(dx))\, n^{1/2}(\widehat{\boldsymbol{\alpha}}_R - \boldsymbol{\alpha}) = -\Lambda_c^{-1}(\boldsymbol{\theta}))\Big[\boldsymbol{S}_\varphi(\boldsymbol{\theta}) + \boldsymbol{G}_c(\boldsymbol{\theta})n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \times \int xg(x)\varphi(G(dx))\Big] + o_p(1).$$

(ii) Moreover, if either $\int xg(x)\varphi(G(dx)) = 0$ *or* $\boldsymbol{G}_c(\boldsymbol{\theta}) = 0$, *then*

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\alpha}}_R - \boldsymbol{\alpha}) \Rightarrow \mathcal{N}_{r_1}\Big[\mathbf{0}, \Lambda_c^{-1}(\boldsymbol{\theta})J_R(\varphi, G)\Big], \tag{41}$$

where $J_R(\varphi, G) = \frac{\int \varphi^2(u)du - (\int \varphi(u)du)^2}{[\int g(x)\varphi(G(dx))]^2}$.

A sufficient condition for $\int xg(x)\varphi(G(dx)) = 0$ is that g is even and φ is skew symmetric, i.e., $\varphi(u) = -\varphi(1-u)$, $\forall u \in [0,1]$. Therefore, in practice, the use of a skew symmetric φ is recommended to ensure that Theorem 2(ii) holds when the innovations are symmetrically distributed. For some models, e.g., in ARLSCH of Example 2, $\boldsymbol{G}_c(\boldsymbol{\theta}) = 0$ when X_0 is symmetrically distributed around zero. However, for Example 1 (Engle's ARCH) and Example 3 (ARTCH), $\boldsymbol{G}_c(\boldsymbol{\theta}) \neq 0$ and the use of a skew symmetric score function is essential.

REMARK 3. The conditions of Theorems 2(ii) and 3(ii) ensure that the preliminary estimator and the scale estimator have no effect on the asymptotics of the final estimator. Using these Theorems, the asymptotic distributions of $\widehat{\boldsymbol{\alpha}}_M$ and $\widehat{\boldsymbol{\alpha}}_R$ are same as those of M- and R-estimator of $\boldsymbol{\alpha}$ for the model (35)

$$\frac{X_i}{\sigma(\boldsymbol{Y}_{i-1}, \boldsymbol{\beta})} = \frac{\mu(\boldsymbol{Y}_{i-1}, \boldsymbol{\alpha})}{\sigma(\boldsymbol{Y}_{i-1}, \boldsymbol{\beta})} + \eta_i,$$

with $\boldsymbol{\beta}$ known. □

REMARK 4. The quasi maximum likelihood estimator $\widehat{\boldsymbol{\alpha}}_{QMLE}$ of $\boldsymbol{\alpha}$ can be defined as a minimizer of

$$\sum_{i=1}^{n}[X_i/\sigma(\boldsymbol{Y}_{i-1}, \widehat{\boldsymbol{\beta}}) - \{\mu(\boldsymbol{Y}_{i-1}, \boldsymbol{a})/\sigma(\boldsymbol{Y}_{i-1}, \widehat{\boldsymbol{\beta}})\}]^2$$

with respect to $\boldsymbol{a}$. Using standard techniques, its asymptotic distribution can be obtained as

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\alpha}}_{QMLE} - \boldsymbol{\alpha}) \Rightarrow \mathcal{N}_{r_1}\Big[\mathbf{0}, \Lambda^{-1}(\boldsymbol{\theta})\Big]. \tag{42}$$

When $\Lambda_c(\boldsymbol{\theta}) = \Lambda(\boldsymbol{\theta})$, we can use (41) and (42) to define the asymptotic relative efficiency (ARE) of an R-estimator based on φ, with respect to the QMLE as $1/J_R(\varphi, G)$.

When $\varphi(u) = u - 1/2$, the ARE of the Wilcoxon R-estimator with respect to the QMLE is $12(\int g^2(x)dx)^2$, which is at least 0.864 for a large class of symmetric standardized error densities g; see, for example, the work done by Lehmann 1983, Section 5.6 for similar result under the location model. In particular, for the standardized normal, logistic, and the double-exponential g, ARE equals $3/\pi(0.955)$, $\pi^2/9(1.10)$, and 1.50, respectively. In a similar fashion, the ARE of the R-estimator based on signed score with respect to the QMLE is $4g^2(0)$, which is at least $1/3$ for symmetric unimodal error densities g (with variance 1); see, for example, Lehmann 1983, Section 5.3 for similar result under the location model. In particular, for the standardized normal, logistic, and double-exponential g, ARE equals $2/\pi(0.637)$, $\pi^2/12(0.82)$, and 2, respectively. □

Example 11 (ARCH Model). In this case with $\boldsymbol{Y}_0' = [1, Z_0^2, \ldots, Z_{1-s}^2]$,

$$\Lambda(\boldsymbol{\theta}) = \boldsymbol{G}(\boldsymbol{\theta}) = E\frac{\boldsymbol{Y}_0\boldsymbol{Y}_0'}{(\boldsymbol{Y}_0'\boldsymbol{\alpha})^2}$$

and

$$\Lambda_c(\boldsymbol{\theta}) = \boldsymbol{G}_c(\boldsymbol{\theta}) = \Lambda(\boldsymbol{\theta}) - \left[E\frac{\boldsymbol{Y}_0}{\boldsymbol{Y}_0'\boldsymbol{\alpha}}\right]\left[E\frac{\boldsymbol{Y}_0}{\boldsymbol{Y}_0'\boldsymbol{\alpha}}\right]'.$$

Clearly, $\boldsymbol{G}(\boldsymbol{\theta})$ and $\boldsymbol{G}_c(\boldsymbol{\theta})$ are nonzero. Hence, from (40), if $\int xg(x)d\psi(x) = 0$, then

$$n^{1/2}(\widehat{\boldsymbol{\alpha}}_M - \boldsymbol{\alpha}) \Rightarrow \mathcal{N}_{s+1}\Big[\mathbf{0}, \Lambda^{-1}(\boldsymbol{\theta})J_M(\psi, G)\Big].$$

From (41), if $\int xg(x)\varphi(G(dx)) = 0$, then

$$n^{1/2}(\widehat{\boldsymbol{\alpha}}_R - \boldsymbol{\alpha}) \Rightarrow \mathcal{N}_{s+1}\Big[\mathbf{0}, \Lambda_c^{-1}(\boldsymbol{\theta})J_R(\varphi, G)\Big].$$

When $E(\varepsilon_1^4) < \infty$ in (30), the asymptotic distribution of the widely used quasi maximum likelihood estimator (QMLE) $\hat{\boldsymbol{\alpha}}_{QMLE}$ is as follows.

$$n^{1/2}(\widehat{\boldsymbol{\alpha}}_{QMLE} - \boldsymbol{\alpha}) \Rightarrow \mathcal{N}_{s+1}\Big[\mathbf{0}, \Lambda^{-1}(\boldsymbol{\theta})Var(\varepsilon_1^2)\Big].$$

Thus, it follows that the asymptotic relative efficiency of an M-estimator $\widehat{\boldsymbol{\alpha}}_M$, relative to the QMLE in Engle's ARCH model is similar to that of the M-estimator relative to the least squared estimator in the one sample location model or in the linear regression model.

Example 12 (ARLSCH Model). In this case with $\boldsymbol{Z}_0 = [1, X_0^2]'$,

$$\Lambda(\boldsymbol{\theta}) = E\left[\frac{X_0^2}{\boldsymbol{\beta}'\boldsymbol{Z}_0}\right],$$

$$\boldsymbol{G}(\boldsymbol{\theta}) = E\left[\frac{X_0\boldsymbol{Z}_0'}{2(\boldsymbol{\beta}'\boldsymbol{Z}_0)^{3/2}}\right],$$

$$\Lambda_c(\boldsymbol{\theta}) = \Lambda(\boldsymbol{\theta}) - \left[E\left\{\frac{X_0}{(\boldsymbol{\beta}'\boldsymbol{Z}_0)^{1/2}}\right\}\right]^2,$$

and

$$G_c(\boldsymbol{\theta}) = G(\boldsymbol{\theta}) - E\Big[\frac{X_0}{(\boldsymbol{\beta}'\boldsymbol{Z}_0)^{1/2}}\Big]E\Big[\frac{Z_0'}{2(\boldsymbol{\beta}'\boldsymbol{Z}_0)}\Big].$$

Note that if X_0 is symmetrically distributed around zero, $G(\boldsymbol{\theta}) = G_c(\boldsymbol{\theta}) = \mathbf{0}$. Therefore, under the existence of moments of appropriate order,

$$n^{1/2}(\widehat{\boldsymbol{\alpha}}_M - \boldsymbol{\alpha}) \Rightarrow \mathcal{N}\Big[0, \Lambda^{-1}(\boldsymbol{\theta})J_M(\psi, G)\Big]$$

and

$$n^{1/2}(\widehat{\boldsymbol{\alpha}}_R - \boldsymbol{\alpha}) \Rightarrow \mathcal{N}\Big[0, \Lambda_c^{-1}(\boldsymbol{\theta})J_R(\varphi, G)\Big].$$

Again, it follows that the asymptotic relative efficiency of the M-estimator corresponding to the score function ψ, relative to the least square estimator, in the above model is similar to that for the one sample location or for the linear regression and autoregressive models; same comment about R-estimator is applicable.

Example 13 (ARTCH Model). In this model, both the mean and the standard deviation are linear in parameters and the expressions for different matrices can be found very easily from (39). We omit details.

Again, a relative efficiency statement similar to the one in the previous two examples holds here as well.

7. Data analysis for the AR–ARCH model

In this section, we first report simulation study verifying the asymptotic distributional results of Section 6.2 and compare the Wilcoxon R-estimator ($\widehat{\boldsymbol{\alpha}}_W$), the R-estimator based on the signed score ($\widehat{\boldsymbol{\alpha}}_S$), and the QMLE ($\widehat{\boldsymbol{\alpha}}_{QMLE}$) at three error densities in terms of their average squared deviations from the true parameter or the estimated mean squared error (MSE). Consequently, the performance of some optimal R-estimators at certain error densities are compared with the Gaussian likelihood–based MLE. Next we consider the monthly log returns of IBM stock from Section 3 and study the robustness of Wilcoxon R-estimator against misspecified form of the heteroscedasticity for this data in comparison with $\hat{\boldsymbol{\alpha}}_{QMLE}$.

7.1. Simulation study

Among many different models, we choose the ARTCH model with $p = s = 1$ and the ARLSCH model with $p = s = 1, r = 2$ when the errors are simulated from the standardized (i) normal (N), (ii) logistic (L), and (iii) double-exponential (D) distribution with specified value of the underlying true parameter $\boldsymbol{\theta}$. To estimate the scale parameters, we use the score function $\kappa(u) = u$. The computations become relatively simpler under such choice of the score function with even closed-form expressions for the scale estimators in the ARTCH model. For each model, we use $r = 100$ replications. For each of the kth replication ($1 \le k \le r$), we generate a sample of size $n = 100$ with parameters $\alpha = 0.1, \beta_1 = 0.2, \beta_2 = 0.3$ for the ARTCH model and $\alpha = 0.1, \beta_0 = 0.2, \beta_1 = 0.3$ for the ARLSCH model and compute (i) the preliminary estimator $\widehat{\alpha}_p$, (ii) the MLE based on the normal distribution $\widehat{\alpha}_{QMLE}$, (iii) the Wilcoxon R-estimator $\widehat{\alpha}_W$ based on the score function $\varphi(u) = u - (1/2)$, and (iv) the R-estimator $\widehat{\alpha}_S$ based on the signed-score function $\varphi(u) = \text{sign}\,\{u - (1/2)\}$. For each estimator (denoted generically by $\widehat{\alpha}(k)$), we also compute $r^{-1}\sum_{k=1}^{r}(\widehat{\alpha}(k) - \alpha)^2$ that is the average (over all replications) squared deviation of the estimate from the true parameter value α, and this is an estimate of mean squared error (MSE) of $\widehat{\alpha}$.

Simulation results and analysis. These are reported in columns (2)–(5) in Table 12 and Table 13. Columns (6) and (8) are obtained from dividing column (5) by columns (3) and (4), respectively, and represent the estimated ARE of $\widehat{\alpha}_W$ and $\widehat{\alpha}_S$ with respect to $\widehat{\alpha}_{QMLE}$ (denoted by $\text{E}(\widehat{\alpha}_W)$ etc.); entries in the bold represent the maximum estimated ARE over different error distributions. Columns (7) and (9) represent the corresponding theoretical ARE of $\widehat{\alpha}_W$ and $\widehat{\alpha}_S$ as explained in Remark 2.2 (denoted by $\text{T}(\widehat{\alpha}_W)$ etc.). For each scenario (corresponding to a particular row in the tables), we have run simulations five times under identical setup and have reported the result of that simulation that has best estimated ARE (in the sense that it is either more than or the closest to the theoretical ARE).

Simulation results as well as several histograms conform with our theoretical finding on the asymptotic normality of the different estimators. In many other simulations not

Table 12
Estimated MSEs and AREs of the different estimators of α (ARTCH model)

g	$\text{MSE}(\widehat{\alpha}_p)$	$\text{MSE}(\widehat{\alpha}_W)$	$\text{MSE}(\widehat{\alpha}_S)$	$\text{MSE}(\widehat{\alpha}_{QMLE})$	$\text{E}(\widehat{\alpha}_W)$	$\text{T}(\widehat{\alpha}_W)$	$\text{E}(\widehat{\alpha}_S)$	$\text{T}(\widehat{\alpha}_S)$
N	0.0545	0.0005	0.0006	0.0005	0.983	0.96	0.940	0.64
L	0.0459	0.0007	0.0007	0.0008	**1.181**	1.1	1.209	0.82
D	0.0416	0.0004	0.0004	0.0007	1.558	1.5	**1.670**	2

Table 13
Estimated MSEs and AREs of the different estimators of α (ARLSCH model)

g	$\text{MSE}(\widehat{\alpha}_p)$	$\text{MSE}(\widehat{\alpha}_W)$	$\text{MSE}(\widehat{\alpha}_S)$	$\text{MSE}(\widehat{\alpha}_{QMLE})$	$\text{E}(\widehat{\alpha}_W)$	$\text{T}(\widehat{\alpha}_W)$	$\text{E}(\widehat{\alpha}_S)$	$\text{T}(\widehat{\alpha}_S)$
N	0.0183	0.0208	0.0291	0.0188	0.903	0.96	0.645	0.64
L	0.0232	0.0136	0.0214	0.0154	**1.139**	1.1	0.721	0.82
D	0.0217	0.0128	0.0133	0.0173	1.354	1.5	**1.300**	2

reported here with different combinations of the underlying parameters, it was observed that the ARE results for $\widehat{\alpha}_W$ and $\widehat{\alpha}_S$ approximately hold even when the models are nonstationary. In general, to a practitioner, we recommend the use of $\widehat{\alpha}_W$ as a good alternative to the QMLE that has high ARE for a wide number of distributions with a "small sacrifice" at the normal distribution. Hence, in the real data examples below, we use only $\widehat{\alpha}_W$ and $\widehat{\alpha}_{QMLE}$ for our analysis.

7.2. *Financial data*

Our main goal is to demonstrate the robustness of the proposed R-estimator $\widehat{\alpha}_R$ against the form of the conditional heteroscedasticity of the model. For illustration, we consider the Wilcoxon R-estimator $\widehat{\alpha}_W$ and demonstrate its robustness by showing that its values corresponding to symmetric AR(1)–ARLSCH and asymmetric AR(1)–TARCH models are close to each other. In fact, they are close to the QMLE estimates of α computed in Tsay (2010) for the symmetric AR(1)–GARCH and asymmetric AR(1)–EGARCH models. At the same time, we further exhibit the extreme sensitivity of the QMLE to the model specification of the conditional heteroscedasticity by noticing that the QMLE estimates for the mean parameter are very different for the AR(1)–ARLSCH and AR(1)–TARCH models. To exhibit the robustness of $\widehat{\alpha}_W$ more convincingly, we need to demonstrate that the R-estimates are close to each other for the symmetric AR(1)–GARCH and asymmetric AR(1)–EGARCH models, but this is beyond the scope of the model (29) and will be taken up in future research.

Tsay (2010, Example 3.4) fitted an AR(1) model with GARCH error to the IBM data to obtain the estimate of the autoregressive parameter as 0.099 with SE 0.037 and the model seemed to be adequate. We use the ARLSCH model to get the preliminary estimate $\widehat{\alpha}_p = 0.10601551$ and the R-estimate $\widehat{\alpha}_W = 0.10864080$ with SE 0.01903097. Therefore, the intercept parameter is close to Tsay's estimate and is significant in accordance with Tsay's result. However, the QMLE for the ARLSCH model is $\widehat{\alpha}_{QMLE} = 0.31733076$ with SE 0.09571206 and is very different than the estimate obtained by Tsay using the QMLE of AR(1)–GARCH model. This shows that $\widehat{\alpha}_W$ is more robust to the specification between the ARCH or GARCH heteroscedasticity than $\widehat{\alpha}_{QMLE}$. Moreover, the estimated ARE of the R-estimator with respect to the QMLE for the ARLSCH model is as high as 25.29363788.

Let $Q^*(k)$ denote the Ljung-Box statistic with lag k for the portmanteau test of the randomness of the residuals $\{\hat{\epsilon}_t\}$ from the ARLSCH model. Using the R-estimate for residuals, the Ljung-Box statistics turn out to be $Q^*(10) = 6.8387$ and $Q^*(20) = 15.0339$, whereas using the QMLE for residuals, $Q^*(10) = 6.9607$ and $Q^*(20) = 14.7694$. Since the Ljung-Box statistics have high p-values, the ARLSCH model seems to be adequate using both the R-estimator and the QMLE.

Next we appeal to the asymmetric feature of this data. Tsay (2010) fitted an AR(1)–EGARCH model to this data to obtain the estimate of the autoregressive parameter as 0.092. Fitting an ARTCH model to this data, we obtain the preliminary estimate 0.10601551 and $\widehat{\alpha}_W = 0.09289947$ with SE 0.14118706. However, the QMLE is very different from this R-estimate and Tsay's comparable estimate with value $\widehat{\alpha}_{QMLE} = 0.41444369$ and SE 0.26747658. Note that the intercept parameter appears to be not significant using both estimates. Next we consider the Ljung-Box statistics for the ARTCH model. With residuals from rank-estimate $Q^*(10) = 7.0857$ and

$Q^*(20) = 31.7230$ while with the QMLE, $Q^*(10) = 7.4309$ and $Q^*(20) = 31.3810$ and the ARTCH model seems to be adequate. This demonstrates, as before, that the R-estimator performs better with model misspecification between the ARTCH and the EGARCH models. Moreover, the estimated ARE of the R-estimator is 3.58906858.

8. Conclusions

In this chapter, we reviewed robust estimation methods of the parameters of the conditional heteroscedastic models such as nonlinear ARCH and GARCH or more generally GJR models. We applied them to financial data sets and used backtesting methods to assess the in-sample and the out-of-sample VaR performance of M-estimators.

From our empirical analysis, it turns out that the VaR estimates are robust against the choice of the functional form of the heteroscedasticity of the model especially on the ground of symmetry and the R-estimates are robust against the misspecified form of the conditional heteroscedasticity. For the GARCH and GJR models, the AQL of the Cauchy and B-estimator was the least among the five M-estimators considered for the cited data sets. The MRB of the QMLE was also found to be higher than other estimators in most of the cases, indicating that the risk estimate of the QMLE was slightly larger than the average of other risk estimates. These findings confirmed the superiority of the Cauchy and B-estimator over the QMLE.

In fact, in many occasions, the QMLE is routinely used without paying attention to the fact that the finite fourth moment assumption is not tenable for that data. In those cases, alternatives to QMLE for which a well-developed asymptotic theory exists provide strong justification for their use.

A number of interesting extensions and questions emerge naturally from this research that need further investigation. For example, it will be of interest to investigate the robustness of R- and M-estimators against the misspecified form of the conditional heteroscedasticity when the mean functional is truly nonlinear in the parameters. As mentioned in Section 7, symmetric AR(p)–GARCH and asymmetric AR(p)–EGARCH models are useful in various data analysis, and it will be interesting to investigate the theoretical and empirical properties of various robust estimators in these setup.

Acknowledgments

I am grateful to the anonymous referee and the Editors for making many constructive comments to improve the original manuscript.

References

Basel Committee on Banking Supervision, 1996. Supervisory Framework for the use of Backtesting in Conjunction with the International Model-based Approach to Market Risk Capital Requirements. BIS, Basel, Switzerland.

Berkes, I., Horvath, L., 2004. The efficiency of the estimators of the parameters in GARCH processes. Ann. Stat. 32, 633–655.

Berkes, I., Horvath, L., Kokoszka, P., 2003. GARCH processes: structure and estimation. Bernoulli 9, 201–228.

Bollerslev, T., 1986. Generalised autoregressive conditional heteroscedasticity. J. Econom. 31, 307–327.
Bose, A., Mukherjee, K., 2003. Estimating the ARCH parameters by solving linear equations. J. Time Ser. Anal. 24, 127–136.
Bougerol, P., Picard, N., 1992. Stationarity of GARCH processes and of some nonnegative time series. J. Econom. 52, 115–127.
Christoffersen, P., 1998. Evaluating interval forecasts. Int. Econ. Rev. 39, 841–862.
Engle, R., 1982. Autoregressive conditional heteroskedasticity and estimates of the variance of UK inflation. Econometrica 50, 987–1008.
Engle, R., Manganelli, S., 2004. CAViaR: Conditional autoregressive value at risk by regression quantiles. J. Bus. Econ. Stat. 22, 367–381.
Fan, J., Qi, L., Xiu, D., 2010. Non-Gaussian Quasi Maximum Likelihood Estimation of GARCH Models. Technical Report. Princeton University. Available at (papers.ssrn.com) (accessed 28.4.2011).
Glosten, L., Jagannathan, R., Runkle, D., 1993. On the relation between the expected value and the volatility on the nominal excess returns on stocks. J. Finance 48, 1779–1801.
Härdle, W., Tsybakov, A., 1997. Local polynomial estimators of the volatility function in nonparametric autoregressive. J. Econom. 81, 223–242.
Hendricks, D., 1996. Evaluation of value-at-Risk models using Historical Data. Federal Reserve Bank of New York, Econ. Pol. Rev. April, 36–69.
Iqbal, F., 2010. Contributions to Conditional Heteroscedastic Models: M-estimation and other Methods. PhD Thesis; Dept. of Mathematics and Statistics, Lancaster University.
Iqbal, F., Mukherjee, K., 2010. M-estimators of some GARCH-type models; computation and application. Stat. Comput. 20, 435–445.
Iqbal, F., Mukherjee, K., in press. A study of Value-at-Risk based on M-estimators of the conditional heteroscedastic models. J. Forecast.
Jorion, P., 2000. Value-at-Risk: The New Benchmark for Managing Financial Risk. McGraw-Hill, New York.
Koul, H., Mukherjee, K., 2002. Some Estimation Procedures in ARCH Models. Technical Report 9-2002. National University of Singapore.
Kupiec, P., 1995. Techniques for verifying the accuracy of risk measurement models. J. Derivatives 3, 73–84.
Lehmann, E., 1983. Theory of Point Estimation. Wiley, New York.
Lopez, J., 1999. Methods for evaluating Value-at-Risk estimates. Fed. Reserve Bank San Francisco Econ. Rev. 2, 3–17.
Mukherjee, K., 2006. Pseudo-likelihood estimation in ARCH model. Can. J. Stat. 34, 341–356.
Mukherjee, K., 2007. Generalized R-estimators under conditional heteroscedasticity. J. Econom. 141, 383–415.
Mukherjee, K., 2008. M-estimation in GARCH model. Econom. Theory 24, 1530–1553.
Nelson, D., 1991. Conditional heteroscedasticity in asset returns; a new approach. Econometrica 59, 347–370.
Peng, L., Yao, Q., 2003. Least absolute deviations estimation for ARCH and GARCH models. Biometrika 90, 967–975.
Rabemananjara, R., Zakoian, J., 1993. Threshold ARCH models and asymmetry in volatility. J. Appl. Econom. 8, 31–49.
Robinson, P., Zaffaroni, P., 2006. Pseudo-maximum-likelihood estimation of ARCH(∞) models. Ann. Stat. 34, 1049–1074.
Tsay, R., 2010. Analysis of Financial Time Series. Wiley, New York.

Part III: High Dimensional Time Series

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00007-7

Functional Time Series

Siegfried Hörmann[1] *and Piotr Kokoszka*[2]

[1]*Department of Mathematics, Université Libre de Bruxelles, Bd du Triomphe, B-1050 Bruxelles, Belgique*
[2]*Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, USA*

Abstract

This chapter is an account of the recent research that deals with curves observed consecutively over time. The curves are viewed in the framework of functional data analysis, that is, each of them is considered as a whole statistical object. We describe the Hilbert space framework within which the mathematical foundations are developed. We then introduce the most popular model for such data, the functional autoregressive process, and discuss its properties. This is followed by the introduction of a general framework that quantifies the temporal dependence of curves. Within this framework, we discuss analogs of central concepts of time series analysis of scalar data, including the definition and the estimation of an analog of the long-run variance.

Keywords: autoregressive process, functional data, prediction, principal components, time series.

1. Introduction

Functional data often arise from measurements obtained by separating an almost continuous time record into natural consecutive intervals, for example, days. Examples include daily curves of financial transaction data and daily patterns of geophysical and environmental data. The functions thus obtained form a time series $\{X_k, k \in \mathbb{Z}\}$, where each X_k is a (random) function $X_k(t)$, $t \in [a, b]$. We refer to such data structures as *functional time series*; examples are given in Section 1.1. A central issue

in the analysis of such data is to take into account the temporal dependence of the observations, i.e., the dependence between events determined by $\{X_k, k \leq m\}$ and $\{X_k, k \geq m + h\}$. Although the literature on scalar and vector time series is huge, there are relatively few contributions dealing with functional time series. The focus of functional data analysis has been mostly on i.i.d. functional observations. Therefore, it is hoped that the present survey will provide an informative account of a useful approach that merges the ideas of time series analysis and functional data analysis.

The monograph of Bosq (2000) studies the theory of linear functional time series, both in Hilbert and Banach spaces, focusing on the functional autoregressive model. For many functional time series, it is, however, not clear what specific model they follow, and for many statistical procedures, it is not necessary to assume a specific model. In such cases, it is important to know what the effect of the dependence on a given procedure is. Is it robust to temporal dependence, or does this type of dependence introduce a serious bias? To answer the questions of this type, it is essential to quantify the notion of temporal dependence. Again, for scalar and vector time series, this question has been approached from many angles, but, except for the linear model of Bosq (2000), for functional time series, no general framework has been available. We present a moment-based notion of weak dependence proposed by Hörmann and Kokoszka (2010).

To make this account as much self-contained as possible, we set in Section 2 the mathematical framework required for this contribution and also report some results for i.i.d. data, to allow for a comparison between results for serially dependent and independent functional data. Next, in Section 3, we introduce the autoregressive model of Bosq (2000) and discuss its applications. In Section 4, we outline the notion of dependence proposed by Hörmann and Kokoszka (2010) and show how it can be applied to the analysis of functional time series. References to related topics are briefly discussed in Section 5.

1.1. Examples of functional time series

The data that motivated the research presented here are of the form $X_k(t)$, $t \in [a, b]$. The interval $[a, b]$ is typically normalized to be a unit interval $[0, 1]$. The treatment of the endpoints depends on the way the data are collected. For intradaily financial transactions data, a is the opening time and b is the closing time of an exchange, for example, the NYSE, so both endpoints are included. Geophysical data are typically of the form $X(u)$, where u is measured at a very fine time grid. After normalizing to the unit interval, the curves are defined as $X_k(t) = X(k + t)$, $0 \leq t < 1$. In both cases, an observation is thus a curve.

Figure 1 shows a reading of a magnetometer over a period of 1 week. A magnetometer is an instrument that measures the three components of the magnetic field at a location where it is placed. There are over 100 magnetic observatories located on the surface of the Earth, and most of them have digital magnetometers. These magnetometers record the strength and direction of the field every 5 seconds, but the magnetic field exists at any moment of time, so it is natural to think of a magnetogram as an

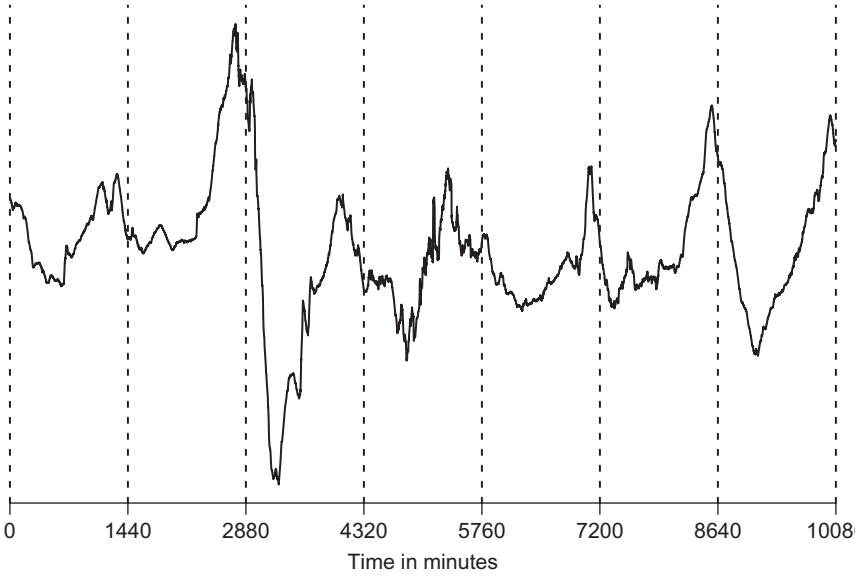

Fig. 1. The horizontal component of the magnetic field measured in 1 min resolution at Honolulu magnetic observatory from January 1, 2001 00:00 UT to January 7, 2001 24:00 UT.

approximation to a continuous record. The raw magnetometer data are cleaned and reported as averages over 1 min intervals. Such averages were used to produce Fig. 1. Thus, $7 \times 24 \times 60 = 10{,}080$ values (of one component of the field) were used to draw Fig. 1. The vertical lines separate days in Universal Time (UT). It is natural to view a curve defined over one UT day as a single observation because one of the main sources influencing the shape of the record is the daily rotation of the Earth. When an observatory faces the Sun, it records the magnetic field generated by wind currents flowing in the ionosphere, which are driven mostly by solar heating. Fig. 1, thus, shows seven consecutive observations of a functional time series.

Examples of data that can be naturally treated as functional also come from financial records. Figure 2 shows 2 consecutive weeks of Microsoft stock prices in 1 min resolution. A great deal of financial research has been done using the closing daily price, i.e., the price in the last transaction of a trading day. However, many assets are traded so frequently that one can practically think of a price curve that is defined at any moment of time. The Microsoft stock is traded several hundred times per minute. The values used to draw the graph in Fig. 2 are the closing prices in 1 min intervals. It is natural to choose one trading day as the underlying time interval. If we do so, Fig. 2 shows 10 consecutive functional observations. In contrast to magnetometer data, the price in the last minute of day k does not have to be close to the price in the first minute of day $k + 1$.

However, we would like to emphasize that functional time series need not arise through the mechanism described above. For example, the Eurodollar curves studied by Kargin and Onatski (2008) and Horváth et al. (2012) are not of this form. A functional time series is a sequence of curves.

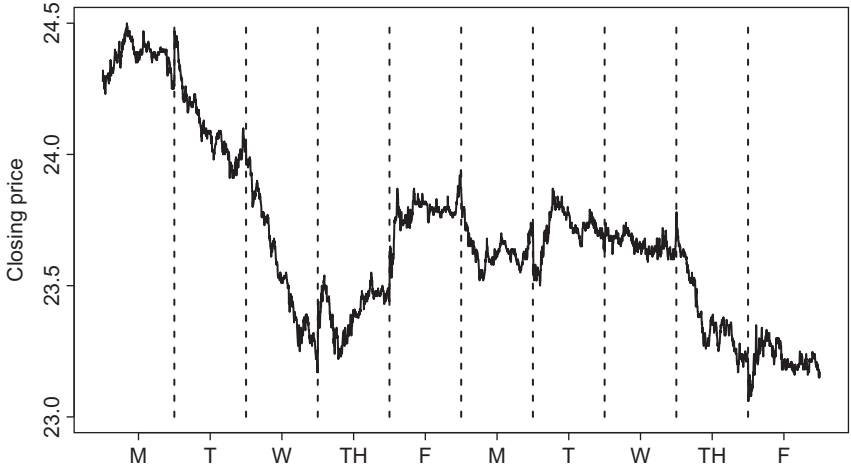

Fig. 2. Microsoft stock prices in 1 min resolution, May 1–5, 8–12, 2006.

2. The Hilbert space model for functional data

It is typically assumed that the observations X_k are elements of a separable Hilbert space H (i.e., a Hilbert space with a countable basis $\{e_k, k \in \mathbb{Z}\}$) with inner product $\langle \cdot, \cdot \rangle$, which generates norm $\| \cdot \|$. This is what we assume in the following. An important example is the Hilbert space $L^2 = L^2([0,1])$ introduced in Section 2.2. Although we formally allow for a general Hilbert space, we call our H–valued data *functional observations*. All random functions are defined on some common probability space $(\Omega, \mathcal{A}, P)$. We say that X is integrable if $E\|X\| < \infty$, and we say it is square integrable if $E\|X\|^2 < \infty$. If $E\|X\|^p < \infty$, $p > 0$, we write $X \in L_H^p = L_H^p(\Omega, \mathcal{A}, P)$. Convergence of $\{X_n\}$ to X in L_H^p means $E\|X_n - X\|^p \to 0$, whereas $\|X_n - X\| \to 0$ almost surely (a.s.) is referred to as almost sure convergence.

In this section, we follow closely the exposition of Bosq (2000). Good references on Hilbert spaces are Riesz and Sz.-Nagy (1990), Akhiezier and Glazman (1993), and Debnath and Mikusinski (2005). An in-depth theory of operators in a Hilbert space is developed by Gohberg et al. (1990).

2.1. Operators

Let $\langle \cdot, \cdot \rangle$ be the inner product in H, which generates the norm $\| \cdot \|$ and denote by $\mathcal{L}$ the space of bounded (continuous) linear operators on H with the norm

$$\|\Psi\|_{\mathcal{L}} = \sup\{\|\Psi(x)\| : \|x\| \leq 1\}.$$

An operator $\Psi \in \mathcal{L}$ is said to be *compact* if there exist two orthonormal bases $\{v_j\}$ and $\{f_j\}$, and a real sequence $\{\lambda_j\}$ converging to zero, such that

$$\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle f_j, \quad x \in H. \tag{1}$$

The λ_j are assumed positive because one can replace f_j by $-f_j$, if needed. Representation (1) is called the *singular value decomposition*. Compact operators are also called *completely continuous* operators.

A compact operator admitting representation (1) is said to be a *Hilbert–Schmidt* operator if $\sum_{j=1}^{\infty} \lambda_j^2 < \infty$. The space $\mathcal{S}$ of Hilbert–Schmidt operators is a separable Hilbert space with the scalar product

$$\langle \Psi_1, \Psi_2 \rangle_{\mathcal{S}} = \sum_{i=1}^{\infty} \langle \Psi_1(e_i), \Psi_2(e_i) \rangle, \tag{2}$$

where $\{e_i\}$ is an arbitrary orthonormal basis, the value of (2) does not depend on it. One can show that $\|\Psi\|_{\mathcal{S}}^2 = \sum_{j \geq 1} \lambda_j^2$ and

$$\|\Psi\|_{\mathcal{L}} \leq \|\Psi\|_{\mathcal{S}}. \tag{3}$$

An operator $\Psi \in \mathcal{L}$ is said to be *symmetric* if

$$\langle \Psi(x), y \rangle = \langle x, \Psi(y) \rangle, \quad x, y \in H,$$

and positive definite if

$$\langle \Psi(x), x \rangle \geq 0, \quad x \in H.$$

(An operator with the last property is sometimes called positive semidefinite, and the term positive definite is used when $\langle \Psi(x), x \rangle > 0$ for $x \neq 0$.)

A symmetric positive definite Hilbert–Schmidt operator Ψ admits the decomposition

$$\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle v_j, \quad x \in H, \tag{4}$$

with orthonormal v_j, which are the eigenfunctions of Ψ, i.e., $\Psi(v_j) = \lambda_j v_j$. The v_j can be extended to a basis by adding a complete orthonormal system in the orthogonal complement of the subspace spanned by the original v_j. The v_j in (4) can thus be assumed to form a basis, but some λ_j may be zero.

2.2. *The space L^2*

The space L^2 is the set of measurable real-valued functions x defined on $[0, 1]$ satisfying $\int_0^1 x^2(t)dt < \infty$. It is a separable Hilbert space with the inner product

$$\langle x, y \rangle = \int x(t)y(t)dt.$$

An integral sign without the limits of integration is meant to denote the integral over the whole interval $[0, 1]$. If $x, y \in L^2$, the equality $x = y$ always means $\int [x(t) - y(t)]^2 dt = 0$.

An important class of operators in L^2 are the integral operators defined by

$$\Psi(x)(t) = \int \psi(t,s)x(s)ds, \quad x \in L^2, \tag{5}$$

with the real kernel $\psi(\cdot,\cdot)$. Such operators are Hilbert–Schmidt if and only if $\iint \psi^2(t,s)dtds < \infty$, in which case

$$\|\Psi\|_S^2 = \iint \psi^2(t,s)dtds. \tag{6}$$

If $\psi(s,t) = \psi(t,s)$ and $\iint \psi(t,s)x(t)x(s)dtds \geq 0$, the integral operator Ψ is symmetric and positive definite, and it follows from (4) that

$$\psi(t,s) = \sum_{j=1}^{\infty} \lambda_j v_j(t)v_j(s) \quad \text{in } L^2([0,1]\times[0,1]). \tag{7}$$

If ψ is continuous, the above expansions holds for all $s,t \in [0,1]$, and the series converges uniformly. This result is known as Mercer's theorem, see e.g., Riesz and Sz.-Nagy (1990).

2.3. *Functional mean and the covariance operator*

Let $X, X_1, X_2, \ldots$ be H–valued random functions. We call X weakly integrable if there is a $\mu \in H$, such that $E\langle X, y\rangle = \langle \mu, y\rangle$ for all $y \in H$. In this case, μ is called the expectation of X, short EX. Some elementary results are (a) EX is unique, (b) integrability implies weak integrability, and (c) $\|EX\| \leq E\|X\|$. In the special case where $H = L^2$, one can show that $\{(EX)(t),\ t \in [0,1]\} = \{E(X(t)),\ t \in [0,1]\}$, i.e., one can obtain the mean function by pointwise evaluation. The expectation commutes with bounded operators, i.e., if $\Psi \in \mathcal{L}$ and X is integrable, then $E\Psi(X) = \Psi(EX)$.

For $X \in L_H^2$, the covariance operator of X is defined by

$$C(y) = E[\langle X - EX, y\rangle (X - EX)], \quad y \in H.$$

The covariance operator C is symmetric and positive definite, with eigenvalues λ_i satisfying

$$\sum_{i=1}^{\infty} \lambda_i = E\|X - EX\|^2 < \infty. \tag{8}$$

Hence C is a symmetric positive definite Hilbert–Schmidt operator admitting representation (4).

The sample mean and the sample covariance operator of $X_1, \ldots, X_N$ are defined as follows:

$$\hat{\mu}_N = \frac{1}{N}\sum_{k=1}^{N} X_k \quad \text{and} \quad \hat{C}_N(y) = \frac{1}{N}\sum_{k=1}^{N} \langle X_k - \hat{\mu}_N, y\rangle (X_k - \hat{\mu}_N), \quad y \in H.$$

The following result implies the consistency of the just defined estimators for i.i.d. samples.

THEOREM 1. *Let $\{X_k\}$ be an H-valued i.i.d. sequence with $EX = \mu$.*

(a) If $X_1 \in L_H^2$ then $E\|\hat{\mu}_N - \mu\|^2 = O(N^{-1})$.
(b) If $X_1 \in L_H^4$ then $E\|\hat{C}\|_S^2 < \infty$ and $E\|C - \hat{C}\|_S^2 = O(N^{-1})$.

In Section 4, we will prove Theorem 1 in a more general framework, namely for a stationary, weakly dependent sequence.

It is easy to see that for $H = L^2$,

$$C(y)(t) = \int c(t,s)y(s)ds, \quad \text{where} \quad c(t,s) = \text{Cov}(X(t), X(s)).$$

The covariance kernel $c(t,s)$ is estimated by

$$\hat{c}(t,s) = \frac{1}{N}\sum_{k=1}^{N}\big(X_k(t) - \hat{\mu}_N(t)\big)\big(X_k(s) - \hat{\mu}_N(s)\big).$$

2.4. *Empirical functional principal components*

Suppose we observe functions $x_1, x_2, \ldots, x_N$. In this section, it is not necessary to view these functions as random, but we can think of them as the observed realizations of random functions in some separable Hilbert space H. We assume that the data have been centered, i.e., $\sum_{i=1}^N x_i = 0$. Fix an integer $p < N$. We think of p as being much smaller than N, typically a single digit number. We want to find an orthonormal basis $u_1, u_2, \ldots, u_p$, such that

$$\hat{S}^2 = \sum_{i=1}^{N}\left\| x_i - \sum_{k=1}^{p}\langle x_i, u_k\rangle u_k \right\|^2$$

is minimized. Once such a basis is found, $\sum_{k=1}^p \langle x_i, u_k\rangle u_k$ is an approximation to x_i. For the p we have chosen, this approximation is uniformly optimal, in the sense of minimizing $\hat{S}^2$. This means that instead of working with infinitely dimensional curves x_i, we can work with p-dimensional vectors

$$\mathbf{x}_i = \big[\langle x_i, u_1\rangle, \langle x_i, u_2\rangle, \ldots, \langle x_i, u_p\rangle\big]^T.$$

This is a central idea of functional data analysis, as to perform any practical calculations, we must reduce the dimension from infinity to a finite number. The functions u_j are collectively called the *optimal empirical orthonormal basis* or *natural orthonormal components*, the words "empirical" and "natural" emphasizing that they are computed directly from the functional data.

The functions $u_1, u_2, \ldots, u_p$ minimizing $\hat{S}^2$ are equal (up to a sign) to the normalized eigenfunctions, $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_p$, of the sample covariance operator,

i.e. $\hat{C}(u_i) = \hat{\lambda}_i u_i$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. The eigenfunctions $\hat{v}_i$ are called the empirical functional principal components (EFPCs) of the data $x_1, x_2, \ldots, x_N$. Thus, the $\hat{v}_i$ are the natural orthonormal components and form the optimal empirical orthonormal basis.

2.5. *Population functional principal components*

Suppose $X_1, X_2, \ldots, X_N$ are zero mean functional observations in H having the same distribution as X. Parallel to Section 2.4, we can ask which orthonormal elements $v_1, \ldots, v_p$ in H minimize

$$E \left\| X - \sum_{i=1}^{p} \langle X, v_i \rangle \, v_i \right\|^2,$$

and the answer is not surprising in the view of Section 2.5. The eigenfunctions v_i of the covariance operator C allow for the "optimal" representation of X. The functional principal components (FPCs) are defined as the eigenfunctions of the covariance operator C of X. The representation

$$X = \sum_{i=1}^{\infty} \langle X, v_i \rangle \, v_i$$

is called the *Karhunen-Loéve* expansion.

The inner product $\langle X_i, v_j \rangle = \int X_i(t) v_j(t) dt$ is called the jth score of X_i. It can be interpreted as the weight of the contribution of the FPC v_j to the curve X_i.

We often estimate the eigenvalues and eigenfunctions of C, but the interpretation of these quantities as parameters, and their estimation, must be approached with care. The eigenvalues must be identifiable, so we must assume that $\lambda_1 > \lambda_2 > \ldots$. In practice, we can estimate only the p largest eigenvalues and assume that $\lambda_1 > \lambda_2 > \ldots > \lambda_p > \lambda_{p+1}$, which implies that the first p eigenvalues are nonzero. The eigenfunctions v_j are defined by $C(v_j) = \lambda_j v_j$, so if v_j is an eigenfunction, then so is $a v_j$, for any nonzero scalar a (by definition, eigenfunctions are nonzero). The v_j are typically normalized, so that $||v_j|| = 1$, but this does not determine the sign of v_j. Thus, if $\hat{v}_j$ is an estimate computed from the data, we can only hope that $\hat{c}_j \hat{v}_j$ is close to v_j, where

$$\hat{c}_j = \text{sign}\left(\langle \hat{v}_j, v_j \rangle\right).$$

Note that $\hat{c}_j$ cannot be computed form the data, so it must be ensured that the statistics we want to work with do not depend on the $\hat{c}_j$.

With these preliminaries in mind, we define the estimated eigenelements by

$$\hat{C}_N(\hat{v}_j) = \hat{\lambda}_j \hat{v}_j, \quad j = 1, 2, \ldots N. \tag{9}$$

The following result, established in the study by Dauxois et al. (1982) and Bosq (2000), is used very often to develop asymptotic arguments.

THEOREM 2. *Assume that the observations $X_1, X_2, \ldots, X_N$ are i.i.d. in H and have the same distribution as X, which is assumed to be in L_H^4 with $EX = 0$. Suppose that*

$$\lambda_1 > \lambda_2 > \ldots > \lambda_d > \lambda_{d+1}. \tag{10}$$

Then, for each $1 \leq j \leq d$,

$$E\left[\|\hat{c}_j \hat{v}_j - v_j\|^2\right] = O(N^{-1}), \quad E\left[|\lambda_j - \hat{\lambda}_j|^2\right] = O(N^{-1}). \tag{11}$$

Theorem 2 implies that, under regularity conditions, the population eigenfunctions can be consistently estimated by the empirical eigenfunctions. If the assumptions do not hold, the direction of the $\hat{v}_k$ may not be close to the v_k. Examples of this type, with many references, are discussed in the study by Johnstone and Lu (2009). These examples show that if the i.i.d. curves are noisy, then (11) fails. Another setting in which (11) may fail is when the curves are sufficiently regular, but the dependence between them is too strong. Such examples are discussed in Hörmann and Kokoszka (2012).

The proof of Theorem 2 is immediate from part (b) of Theorem 1, and Lemmas 1 and 2, which we will also use in Section 4. These two Lemmas appear, in a slightly more specialized form, as Lemmas 4.2 and 4.3 of Bosq (2000). Lemma 1 is proven in Section 6.1 of Gohberg et al. (1990), see their Corollary 1.6 on p. 99, whereas Lemma 2 is established in the study by Horváth and Kokoszka (2012). To formulate Lemmas 1 and 2, we consider two compact operators $C, K \in \mathcal{L}$ with singular value decompositions

$$C(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle f_j, \quad K(x) = \sum_{j=1}^{\infty} \gamma_j \langle x, u_j \rangle g_j. \tag{12}$$

LEMMA 1. *Suppose $C, K \in \mathcal{L}$ are two compact operators with singular value decompositions (12). Then, for each $j \geq 1$, $|\gamma_j - \lambda_j| \leq \|K - C\|_{\mathcal{L}}$.*

We now define

$$v_j' = c_j v_j, \quad c_j = \operatorname{sign}\left(\langle u_j, v_j \rangle\right).$$

LEMMA 2. *Suppose $C, K \in \mathcal{L}$ are two compact operators with singular value decompositions (12). If C is symmetric, $f_j = v_j$ in (12), and its eigenvalues satisfy (10), then*

$$\|u_j - v_j'\| \leq \frac{2\sqrt{2}}{\alpha_j} \|K - C\|_{\mathcal{L}}, \quad 1 \leq j \leq d,$$

where $\alpha_1 = \lambda_1 - \lambda_2$ and $\alpha_j = \min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})$, $2 \leq j \leq d$.

We note that if C is a covariance operator, then it satisfies the conditions imposed on C in Lemma 2. The v_j are then the eigenfunctions of C. Because these eigenfunctions are determined only up to a sign, it is necessary to introduce the functions v_j'.

This section has merely set out the fundamental definitions and properties. Interpretation and estimation of the functional principal components has been a subject of extensive research, in which concepts of smoothing and regularization play a major role, see Chapters 8, 9, 10 of the study by Ramsay and Silverman (2005).

3. Functional autoregressive model

The theory of autoregressive and more general linear processes in Hilbert and Banach spaces is developed in the monograph of Bosq (2000), on which Sections 3.1 and 3.2 are based and on which we also refer to for the proofs. We present only a few selected results that provide an introduction to the central ideas. Section 3.3 is devoted to prediction by means of the functional autoregressive (FAR) process. To lighten the notation, we set in this chapter, $\|\cdot\|_{\mathcal{L}} = \|\cdot\|$.

3.1. Existence

We say that a sequence $\{X_n,\ -\infty < n < \infty\}$ of mean zero functions in H follows a functional AR(1) model if

$$X_n = \Psi(X_{n-1}) + \varepsilon_n, \tag{13}$$

where $\Psi \in \mathcal{L}$ and $\{\varepsilon_n,\ -\infty < n < \infty\}$ is a sequence of i.i.d. mean zero errors in H satisfying $E\|\varepsilon_n\|^2 < \infty$.

The above definition defines a somewhat narrower class of processes than that considered by Bosq (2000) who does not assume that the $\{\varepsilon_n\}$ are i.i.d., but rather that they are uncorrelated in an appropriate Hilbert space sense, see his Definitions 3.1 and 3.2. The theory of estimation for the process (13) is, however, developed only under the assumption that the errors are i.i.d.

Scalar AR(1) equations, $X_n = \psi X_{n-1} + \varepsilon_n$, admit the unique causal solution $X_n = \sum_{j=0}^{\infty} \psi^j \varepsilon_{n-j}$, if $|\psi| < 1$. Our goal in this section is to state a condition analogous to $|\psi| < 1$ for functional AR(1) equations (13). We begin with the following Lemma:

LEMMA 3. *For any $\Psi \in \mathcal{L}$, the following two conditions are equivalent:*

C0: The exists an integer j_0 such that $\|\Psi^{j_0}\| < 1$.

C1: There exist $a > 0$ and $0 < b < 1$ such that for every $j \geq 0$, $\|\Psi^j\| \leq ab^j$.

Note that condition C0 is weaker than the condition $\|\Psi\| < 1$; in the scalar case, these two conditions are clearly equivalent. Nevertheless, C1 is a sufficiently strong condition to ensure the convergence of the series $\sum_j \Psi^j(\varepsilon_{n-j})$, and the existence of a stationary causal solution to functional AR(1) equations, as stated in Theorem 3.

Note that (13) can be viewed as an iterated random function system, see Diaconis and Freeman (1999) and Wu and Shao (2004). Condition C1 then refers to a geometric contraction property needed to obtain stationary solutions for such processes. Because iterated random function systems have been studied on general metric spaces, we could

use this methodology to investigate extensions of the functional AR process to nonlinear functional Markov processes of the form $X_t = \Psi_{\varepsilon_t}(X_{t-1})$.

THEOREM 3. *If condition C0 holds, then there is a unique strictly stationary causal solution to (13). This solution is given by*

$$X_n = \sum_{j=0}^{\infty} \Psi^j(\varepsilon_{n-j}). \tag{14}$$

The series converges almost surely and in L_H^2.

Example 1. Consider an integral Hilbert–Schmidt operator on L^2 defined by (5), which satisfies

$$\iint \psi^2(t,s)dtds < 1. \tag{15}$$

Recall from Section 2.2 that the left-hand side of (15) is equal to $\|\Psi\|_{\mathcal{S}}^2$. Since $\|\Psi\| \leq \|\Psi\|_{\mathcal{S}}$, we see that (15) implies condition C0 of Lemma 3 with $j_0 = 1$.

3.2. Estimation

This section is devoted to the estimation of the autoregressive operator Ψ, but first we state a theorem on the convergence of the EFPCs and the corresponding eigenvalues, which follows from Example 2, Theorem 7, and Lemma 3. In essence, Theorem 4 states that bounds (11) also hold if the X_n follow an FAR(1) model.

THEOREM 4. *Suppose the operator* Ψ *in (13) satisfies condition C0 of Lemma 3, and the solution* $\{X_n\}$ *satisfies* $E\|X_0\|^4 < \infty$. *If (10) holds, then, for each* $1 \leq j \leq d$, *relations (11) hold.*

We now turn to the estimation of the autoregressive operator Ψ. It is instructive to focus first on the univariate case $X_n = \psi X_{n-1} + \varepsilon_n$, in which all quantities are scalars. We assume that $|\psi| < 1$, so that there is a stationary solution, such that ε_n is independent of X_{n-1}. Then, by multiplying the AR(1) equation by X_{n-1} and taking the expectation, we obtain $\gamma_1 = \psi\gamma_0$, where $\gamma_k = E[X_n X_{n+k}] = \mathrm{Cov}(X_n, X_{n+k})$. The autocovariances γ_k are estimated by the sample autocovariances $\hat{\gamma}_k$, so the usual estimator of ψ is $\hat{\psi} = \hat{\gamma}_1/\hat{\gamma}_0$. This estimator is optimal in many ways, see Chapter 8 of Brockwell and Davis (1991), and the approach outlined above, known as the Yule-Walker estimation, works for higher order and multivariate autoregressive processes. To apply this technique to the functional model, note that by (13), under condition C0 of Lemma 3,

$$E\left[\langle X_n, x\rangle X_{n-1}\right] = E\left[\langle \Psi(X_{n-1}), x\rangle X_{n-1}\right], \quad x \in H.$$

Define the lag-1 autocovariance operator by

$$C_1(x) = E[\langle X_n, x\rangle X_{n+1}]$$

and denote with superscript T, the adjoint operator. Then, $C_1^T = C\Psi^T$ because, by a direct verification, $C_1^T = E\left[\langle X_n, x\rangle\, X_{n-1}\right]$, i.e.,

$$C_1 = \Psi C. \tag{16}$$

The above identity is analogous to the scalar case, so we would like to obtain an estimate of Ψ by using a finite sample version of the relation $\Psi = C_1 C^{-1}$. However, the operator C does not have a bounded inverse on the whole of H. To see it, recall that C admits representation (4), which implies that $C^{-1}(C(x)) = x$, where

$$C^{-1}(y) = \sum_{j=1}^{\infty} \lambda_j^{-1} \left\langle y, v_j \right\rangle v_j.$$

The operator C^{-1} is defined if all λ_j are positive. Because $\|C^{-1}(v_n)\| = \lambda_n^{-1} \to \infty$, as $n \to \infty$, it is unbounded. This makes it difficult to estimate the bounded operator Ψ using the relation $\Psi = C_1 C^{-1}$. A practical solution is to use only the first p, the most important EFPC's $\hat{v}_j$, and to define

$$\widehat{IC}_p(x) = \sum_{j=1}^{p} \hat{\lambda}_j^{-1} \left\langle x, \hat{v}_j \right\rangle \hat{v}_j.$$

The operator $\widehat{IC}_p$ is defined on the whole of L^2, and it is bounded if $\hat{\lambda}_j > 0$ for $j \le p$. By judiciously choosing p, we find a balance between retaining the relevant information in the sample and the danger of working with the reciprocals of small eigenvalues $\hat{\lambda}_j$. To derive a computable estimator of Ψ, we use an empirical version of (16). Because C_1 is estimated by

$$\widehat{C}_1(x) = \frac{1}{N-1} \sum_{k=1}^{N-1} \langle X_k, x\rangle\, X_{k+1},$$

we obtain, for any $x \in H$,

$$\begin{aligned}
\widehat{C}_1 \widehat{IC}_p(x) &= \widehat{C}_1 \left(\sum_{j=1}^{p} \hat{\lambda}_j^{-1} \left\langle x, \hat{v}_j \right\rangle \hat{v}_j \right) \\
&= \frac{1}{N-1} \sum_{k=1}^{N-1} \left\langle X_k, \sum_{j=1}^{p} \hat{\lambda}_j^{-1} \left\langle x, \hat{v}_j \right\rangle \hat{v}_j \right\rangle X_{k+1} \\
&= \frac{1}{N-1} \sum_{k=1}^{N-1} \sum_{j=1}^{p} \hat{\lambda}_j^{-1} \left\langle x, \hat{v}_j \right\rangle \left\langle X_k, \hat{v}_j \right\rangle X_{k+1}.
\end{aligned}$$

The estimator $\widehat{C}_1\widehat{IC}_p$ can be used in principle, but typically an additional smoothing step is introduced by using the approximation $X_{k+1} \approx \sum_{i=1}^{p} \langle X_{k+1}, \hat{v}_i \rangle \hat{v}_i$. This leads to the estimator

$$\widehat{\Psi}_p(x) = \frac{1}{N-1} \sum_{k=1}^{N-1} \sum_{j=1}^{p} \sum_{i=1}^{p} \hat{\lambda}_j^{-1} \langle x, \hat{v}_j \rangle \langle X_k, \hat{v}_j \rangle \langle X_{k+1}, \hat{v}_i \rangle \hat{v}_i. \tag{17}$$

To establish the consistency of this estimator, it must be assumed that $p = p_N$ is a function of the sample size N. Theorem 8.7 of Bosq (2000) then establishes sufficient conditions for $\|\widehat{\Psi}_p - \Psi\|$ to tend to zero. They are technical, but, intuitively, they mean that the λ_j and the distances between them cannot tend to zero too fast.

If $H = L^2$, the estimator (17) is a kernel operator with the kernel

$$\hat{\psi}_p(t,s) = \frac{1}{N-1} \sum_{k=1}^{N-1} \sum_{j=1}^{p} \sum_{i=1}^{p} \hat{\lambda}_j^{-1} \langle X_k, \hat{v}_j \rangle \langle X_{k+1}, \hat{v}_i \rangle \hat{v}_j(s) \hat{v}_i(t). \tag{18}$$

This is verified by noting that

$$\widehat{\Psi}_p(x)(t) = \int \hat{\psi}_p(t,s) x(s) ds.$$

All quantities at the right-hand side of (18) are available as output of the `R` function `pca.fd`, so this estimator is very easy to compute. Kokoszka and Zhang (2010) conducted a number of numerical experiments to determine how close the estimated surface $\hat{\psi}_p(t,s)$ is to the surface $\psi(t,s)$ used to simulate an FAR(1) process. Broadly speaking, for $N \leq 100$, the discrepancies are very large, both in magnitude and in shape. This is illustrated in Fig. 3, which shows the Gaussian kernel $\psi(t,s) = \alpha \exp\left\{-(t^2+s^2)/2\right\}$, with α chosen, so that the Hilbert–Schmidt norm of ψ is $1/2$, and three estimates that use $p = 2, 3, 4$. The innovations ε_n were generated as Brownian bridges. Such discrepancies are observed for other kernels and other innovation processes as well. Moreover, by any reasonable measure of a distance between two surfaces, the distance between ψ and $\hat{\psi}_p$ increases as p increases. This is counterintuitive because by using more EFPC's $\hat{v}_j$, we would expect the approximation (18) to improve. For the FAR(1) used to produce Fig. 3, the sums $\sum_{j=1}^{p} \hat{\lambda}_j$ explain, respectively, 74, 83, and 87% of the variance for $p = 2, 3$, and 4, but (for the series length $N = 100$) the absolute deviation distances between ψ and $\hat{\psi}_p$ are 0.40, 0.44, and 0.55. The same pattern is observed for the RMSE distance $\|\hat{\psi} - \psi\|_{\mathcal{S}}$ and the relative absolute distance. As N increases, these distances decrease, but their tendency to increase with p remains. This problem is partially due to the fact that for many FAR(1) models, the estimated eigenvalues $\hat{\lambda}_j$ are very small, except $\hat{\lambda}_1$ and $\hat{\lambda}_2$, and so a small error in their estimation translates to a large error in the reciprocals $\hat{\lambda}_j^{-1}$ appearing in (18). Kokoszka and Zhang (2010) show that this problem can be alleviated to some extent by adding a positive baseline to the $\hat{\lambda}_j$. However, as we will see in Section 3.3, precise estimation of the kernel ψ is not necessary to obtain satisfactory predictions.

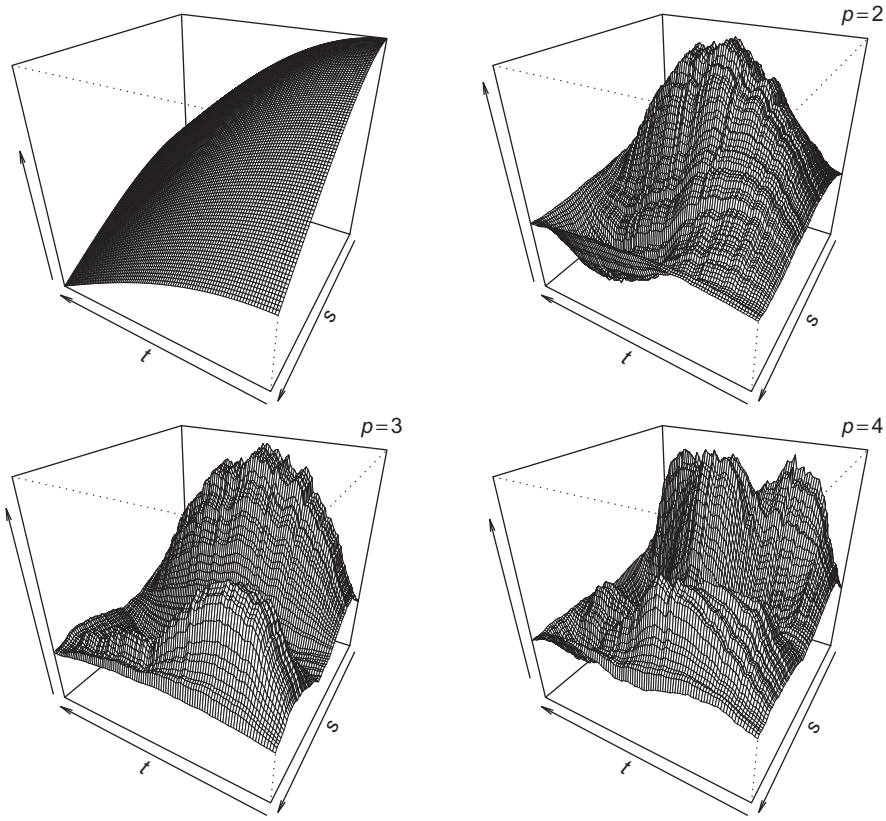

Fig. 3. The kernel surface $\psi(t,s)$ (top left) and its estimates $\hat{\psi}_p(t,s)$ for $p = 2, 3, 4$.

3.3. Prediction

In this section, we discuss some properties of forecasts with the FAR(1) model. Besse et al. (2000) apply several prediction methods, including traditional (nonfunctional) methods, to functional time series derived from real geophysical data. Their conclusion is that the method that we call below *Estimated Kernel* performs best. A different approach to the prediction of functional data was proposed by Antoniadis et al. (2006). In this section, we mostly report the findings of Didericksen et al. (in press), whose simulation study includes a new method proposed by Kargin and Onatski (2008), which we call below *Predictive Factors* and which seeks to replace the FPC's by directions that are most relevant for predictions.

We begin by describing the prediction methods we compare. This is followed by the discussion of their finite sample properties.

3.3.1. Estimated Kernel (EK)

This method uses estimator (18). The predictions are calculated as

$$\hat{X}_{n+1}(t) = \int \hat{\psi}_p(t,s)X_n(s)ds = \sum_{i=1}^{p}\left(\sum_{j=1}^{p}\hat{\psi}_{ij}\langle X_n, \hat{v}_j\rangle\right)\hat{v}_i(t), \tag{19}$$

where

$$\hat{\psi}_{ij} = \hat{\lambda}_j^{-1}(N-1)^{-1}\sum_{n=1}^{N-1}\langle X_n, \hat{v}_j\rangle\langle X_{n+1}, \hat{v}_i\rangle. \tag{20}$$

There are several variants of this method, which depend on where and what kind of smoothing is applied. In our implementation, all curves are converted to functional objects in `R` using 99 Fourier basis functions. The same minimal smoothing is used for the Predictive Factors method.

3.3.2. Predictive Factors (PF)

Estimator (18) is not directly justified by the problem of prediction, it is based on FPCs, which may focus on the features of the data that are not relevant to prediction. An approach known as predictive factors may (potentially) be better suited for forecasting. It finds directions most relevant to prediction, rather than explaining the variability, as the FPCs do. Roughly speaking, it focuses on the optimal expansion of $\Psi(X_n)$, which is, theoretically, the best predictor of X_{n+1}, rather than the optimal expansion of X_n. Because Ψ is unknown, Kargin and Onatski (2008) developed a way of approximating such an expansion in finite samples. We describe only the general idea, as theoretical arguments developed by Kargin and Onatski (2008) are quite complex. As we will see, the PF method does not offer an advantage in finite samples, so we do not need to describe all details here.

Denote by $\mathcal{R}_k$ the set of all rank k operators, i.e., those operators that map L^2 into a subspace of dimension k. The goal is to find $A \in \mathcal{R}_k$ that minimizes $E||X_{n+1} - A(X_n)||^2$. To find a computable approximation to the operator A, a parameter $\alpha > 0$ must be introduced. Following the recommendation of Kargin and Onatski (2008), we used $\alpha = 0.75$. The prediction is computed as

$$\hat{X}_{n+1} = \sum_{i=1}^{k}\left\langle X_n, \hat{b}_{\alpha,i}\right\rangle \hat{C}_1(\hat{b}_{\alpha,i}),$$

where

$$\hat{b}_{\alpha,i} = \sum_{j=1}^{p}\hat{\lambda}_j^{-1/2}\left\langle \hat{x}_{\alpha,i}, \hat{v}_j\right\rangle \hat{v}_j + \alpha\hat{x}_{\alpha,i}.$$

The vectors $\hat{x}_{\alpha,i}$ are linear combinations of the EFPC $\hat{v}_i$, $1 \le i \le k$ and are approximations to the eigenfunctions of the operator Φ defined by the polar decomposition $\Psi C^{1/2} = U\Phi^{1/2}$, where C is the covariance operator of X_1 and U is a unitary operator. The operator $\hat{C}_1$ is the lag-1 autocovariance operator defined by

$$\hat{C}_1(x) = \frac{1}{N-1}\sum_{i=1}^{N-1}\langle X_i, x\rangle\, X_{i+1}, \quad x \in L^2.$$

The method depends on a selection of p and k. We selected p by the cumulative variance method and set $k = p$.

We selected five prediction methods for comparison, two of which do not use the autoregressive structure. To obtain further insights, we also included the errors obtained by assuming perfect knowledge of the operator Ψ. For ease of reference, we now describe these methods and introduce some convenient notation.

MP (Mean Prediction): We set $\hat{X}_{n+1}(t) = 0$. Because the simulated curves have mean zero at every t, this corresponds to using the mean function as a predictor. This predictor is optimal if the data are uncorrelated.

NP (Naive Prediction): We set $\hat{X}_{n+1} = X_n$. This method does not attempt to model temporal dependence. It is included to see how much can be gained by utilizing the autoregressive structure of the data.

EX (Exact): We set $\hat{X}_{n+1} = \Psi(X_n)$. This is not really a prediction method because the autoregressive operator Ψ is unknown. It is included to see whether poor predictions might be due to poor estimation of Ψ (cf. Section 3.2).

EK (Estimated Kernel): This method is described above.

EKI (Estimated Kernel Improved): This is method EK, but the $\hat{\lambda}_i$ in (20) are replaced by $\hat{\lambda}_i + \hat{b}$, as described in Section 3.2.

PF (Predictive Factors): This method is described above.

Didericksen et al. (2011) studied the errors E_n and R_n, $N - 50 < n < N$, defined by

$$E_n = \sqrt{\int_0^1 \left(X_n(t) - \hat{X}_n(t)\right)^2 dt} \quad \text{and} \quad R_n = \int_0^1 \left|X_n(t) - \hat{X}_n(t)\right| dt.$$

for $N = 50, 100, 200$, and $||\Psi||_{\mathcal{S}} = 0.5, 0.8$. They considered several kernels and innovation processes, including smooth errors obtained as a sum of two trigonometric functions, irregular errors generated as Brownian bridges, and intermediate errors obtained by adding small multiples of Brownian bridges to smooth innovations. Examples of boxplots are shown in Figs. 4 and 5. In addition to boxplots, Didericksen et al. (2011) reported the averages of the E_n and R_n, $N - 50 < n < N$, and the standard errors of these averages, which allow to assess whether the differences in the performance of the predictors are statistically significant. Their conclusions can be summarized as follows:

1. Taking the autoregressive structure into account reduces prediction errors, but, in some settings, this reduction is not statistically significant relative to method MP, especially if $\|\Psi\| = 0.5$. Generally if $\|\Psi\| = 0.8$, using the autoregressive structure significantly and visibly improves the predictions.
2. None of the methods EX, EK, and EKI uniformly dominates the other. In most cases, EK method is the best, or at least as good at the others.
3. In some cases, PF method performs visibly worse than the other methods, but always better than NP.
4. Using the improved estimation described in Section 3.2 does not generally reduce prediction errors.

Didericksen et al. (in press) also applied all prediction methods to mean corrected precipitation data studied by Besse et al. (2000). For this data set, the averages of the E_n

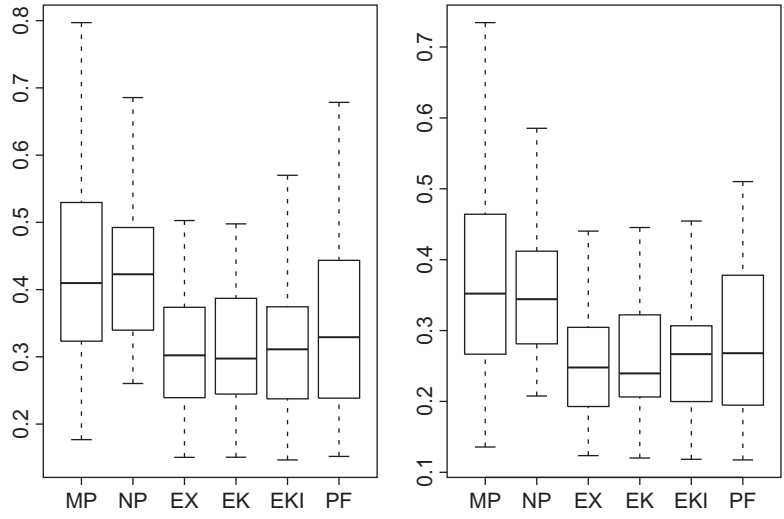

Fig. 4. Boxplots of the prediction errors E_n (left) and R_n (right); Brownian bridge innovations, $\psi(t,s) = Ct$, $N = 100$, $p = 3$, $||\Psi|| = 0.5$.

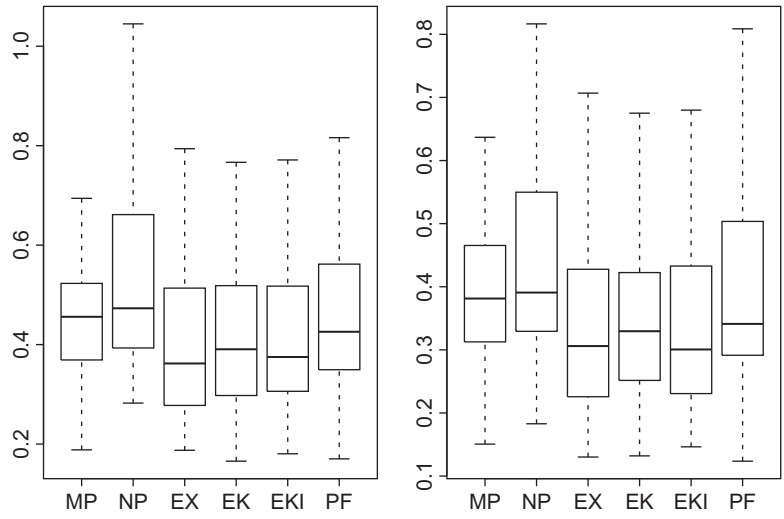

Fig. 5. Boxplots of the prediction errors E_n (left) and R_n (right); Brownian bridge innovations, $\psi(t,s) = Ct$, $N = 100$, $p = 3$, $||\Psi|| = 0.8$.

and the R_n are not significantly different between the first five methods, PF method performs significantly worse than the others. We should point out that PF method depends on the choice of the parameters α and k. It is possible that its performance can be improved by better tuning these parameters. On the other hand, our simulations show that EK method essentially reaches the limit of what is possible, it is comparable to the theoretically perfect EX method. While taking into account the autoregressive structure of the observations does reduce prediction errors, many prediction errors are comparable to those of the trivial MP method. To analyze this observation further, we present in

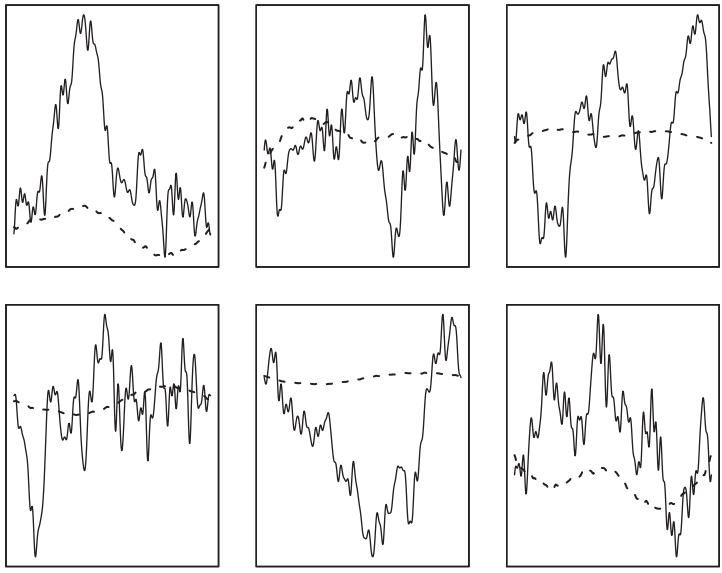

Fig. 6. Six consecutive trajectories of the FAR(1) process with Gaussian kernel, $||\Psi|| = 0.5$, and Brownian bridge innovations. Dashed lines show EK predictions with $p = 3$.

Fig. 6 six consecutive trajectories of a FAR(1) process with $||\Psi|| = 0.5$, and Brownian bridge innovations, together with EK predictions. Predictions obtained with other non-trivial methods look similar. We see that the predictions look much smoother than the observations, and their range is much smaller. If the innovations ε_n are smooth, the observations are also smooth, but the predicted curves have a visibly smaller range than the observations. The smoothness of the predicted curves follows from representation (19), which shows that each predictor is a linear combination of a few EFPCs, which are smooth curves themselves. The smaller range of the the predictors is not peculiar to functional data but is enhanced in the functional setting. For a mean zero scalar AR(1) process $X_n = \psi X_{n-1} + \varepsilon_n$, we have $\mathrm{Var}(X_n) = \psi^2 \mathrm{Var}(X_{n-1}) + \mathrm{Var}(\varepsilon_n)$, so the variance of the predictor $\hat{\psi} X_{n-1}$ is about ψ^{-2} times smaller than the variance of X_n. In the functional setting, the variance of $\hat{X}_n(t)$ is close to $\mathrm{Var}[\int \psi(t,s) X_n(s) ds]$. If the kernel ψ admits the decomposition $\psi(t,s) = \psi_1(t)\psi_2(s)$, as all the kernels we use do, then

$$\mathrm{Var}\left[\hat{X}_n(t)\right] \approx \psi_1^2(t) \mathrm{Var}\left[\int_0^1 \psi_2(s) X_{n-1}(s) ds\right].$$

If the function ψ_1 is small for some values of $t \in [0,1]$, it will automatically drive down the predictions. If ψ_2 is small for some $s \in [0,1]$, it will reduce the integral $\int_0^1 \psi_2(s) X_{n-1}(s) ds$. The estimated kernels do not admit a factorization of this type, but they are always weighted sums of products of orthonormal functions (the EFPC's $\hat{v}_k$). A conclusion of this discussion is that the predicted curves will in general look

smoother and "smaller" than the data. This somewhat disappointing performance is, however, not due to the poor prediction methods but due to a natural limit of predictive power of the FAR(1) model; the curves $\Psi(X_n)$ share the general properties of the curves $\hat{\Psi}(X_n)$, no matter how Ψ is estimated.

4. Weakly dependent functional time series

What distinguishes time series analysis from other fields of statistics is attention to temporal dependence of the data. In this section, we describe a general framework that accommodates the temporal dependence of functional time series and illustrate it with several examples.

4.1. Approximable functional sequences

The notion of weak dependence has been formalized in many ways. Perhaps the most popular are various mixing conditions, see Doukhan (1994), and Bradley (2007), but in recent years, several other approaches have also been introduced, see Doukhan and Louhichi (1999) and Wu (2005, 2007), among others. In time series analysis, moment-based measures of dependence, most notably autocorrelations and cumulants, have gained wide acceptance. The measure we consider below is a moment type quantity, but it is also related to the mixing conditions as it considers σ–algebras m time units apart, with m tending to infinity. A most direct relaxation of independence is the m-dependence. Suppose $\{X_n\}$ is a sequence of random elements taking values in a measurable space S. Denote by $\mathcal{F}_k^- = \sigma\{\ldots X_{k-2}, X_{k-1}, X_k\}$ and $\mathcal{F}_k^+ = \sigma\{X_k, X_{k+1}, X_{k+2}, \ldots\}$, the σ-algebras generated by the observations up to time k and after time k, respectively. Then the sequence $\{X_n\}$ is said to be m-dependent if for any k, the σ-algebras $\mathcal{F}_k^-$ and $\mathcal{F}_{k+m}^+$ are independent. Most time series models are not m-dependent. Rather, various measures of dependence decay sufficiently fast, as the distance m between the σ-algebras $\mathcal{F}_k^-$ and $\mathcal{F}_{k+m}^+$ increase. However, m-dependence can be used as a tool to study the properties of many nonlinear sequences, see e.g., Hörmann (2008) and Berkes et al. (2011) for recent applications. The general idea is to approximate $\{X_n, n \in \mathbb{Z}\}$ by m-dependent processes $\{X_n^{(m)}, n \in \mathbb{Z}\}$, $m \geq 1$. The goal is to establish that for every n, the sequence $\{X_n^{(m)}, m \geq 1\}$ converges in some sense to X_n, if we let $m \to \infty$. If the convergence is fast enough, then one can obtain the limiting behavior of the original process from corresponding results for m-dependent sequences. Definition 1 formalizes this idea and sets up the necessary framework for the construction of such m-dependent approximation sequences. The idea of approximating scalar sequences by m-dependent nonlinear moving averages appears already in Section 21 of Billingsley (1968), and it was developed in several direction by Pötscher and Prucha (1997). A version of Definition 1 for vector-valued processes was used in the study by Aue et al. (2009).

For $X \in L_H^p$, we define

$$\nu_p(X) = \left(E\|X\|^p\right)^{1/p} < \infty. \tag{21}$$

DEFINITION 1. A sequence $\{X_n\} \in L_H^p$ is called L^p–m–*approximable* if each X_n admits the representation

$$X_n = f(\varepsilon_n, \varepsilon_{n-1}, \ldots), \tag{22}$$

where the ε_i are i.i.d. elements taking values in a measurable space S, and f is a measurable function $f\colon S^\infty \to H$. Moreover, we assume that if $\{\varepsilon_i'\}$ is an independent copy of $\{\varepsilon_i\}$ defined on the same probability space, then letting

$$X_n^{(m)} = f(\varepsilon_n, \varepsilon_{n-1}, \ldots, \varepsilon_{n-m+1}, \varepsilon_{n-m}', \varepsilon_{n-m-1}', \ldots), \tag{23}$$

we have

$$\sum_{m=1}^{\infty} \nu_p\big(X_m - X_m^{(m)}\big) < \infty. \tag{24}$$

The applicability of Definition 1 was demonstrated by Hörmann and Kokoszka (2010) for several linear and nonlinear functional time series. The variables ε_n are typically model errors. The general idea is that in the nonlinear moving average representation (22), the impact on X_n of the ε_{n-m} becomes so small as $m \to \infty$ that they can be replaced by different errors. We illustrate it for the FAR(1) model.

Example 2 (Functional autoregressive process). Let $\{X_n, n \in \mathbb{Z}\}$ be a functional AR(1) model as given as in (13), with $\|\Psi\| < 1$. As we have obtained in Theorem 3, the AR(1) sequence admits the expansion $X_n = \sum_{j=0}^{\infty} \Psi^j(\varepsilon_{n-j})$, where Ψ^j is the j-th iterate of the operator Ψ. We thus set $X_n^{(m)} = \sum_{j=0}^{m-1} \Psi^j(\varepsilon_{n-j}) + \sum_{j=m}^{\infty} \Psi^j(\varepsilon_{n-j}')$. It is easy to verify that for every A in $\mathcal{L}$, $\nu_p(A(Y)) \le \|A\|\, \nu_p(Y)$. Because $X_m - X_m^{(m)} = \sum_{j=m}^{\infty} \big(\Psi^j(\varepsilon_{m-j}) - \Psi^j(\varepsilon_{m-j}')\big)$, it follows that $\nu_p(X_m - X_m^{(m)}) \le 2\sum_{j=m}^{\infty} \|\Psi\|^j\, \nu_p(\varepsilon_0) = O(1)\nu_p(\varepsilon_0)\|\Psi\|^m$. By assumption $\nu_2(\varepsilon_0) < \infty$ and therefore $\sum_{m=1}^{\infty} \nu_2(X_m - X_m^{(m)}) < \infty$, so condition (24) holds with $p \ge 2$, as long as $\nu_p(\varepsilon_0) < \infty$.

4.2. Estimation of the mean function and the FPCs

With the notion of weak dependence just defined at hand, we can obtain analogs of Theorems 1 and 2 for time series. We include the proofs because they illustrate the ease with which the condition of m–approximability is applied. We let $\hat{\mu}_N$ and $\hat{C}_N$ be defined as in Section 2.3.

THEOREM 5. *Assume that $\{X_k\}$ is an H-valued L^2–m–approximable process with $EX = \mu$. Then $E\|\hat{\mu}_N - \mu\|^2 = O(N^{-1})$.*

PROOF. Observe that for any $h > 0$ we have

$$X_0 = f(\varepsilon_0, \varepsilon_{-1}, \ldots), \quad X_h^{(h)} = f^{(h)}(\varepsilon_h, \varepsilon_{h-1}, \ldots, \varepsilon_1, \varepsilon_0^{(h)}, \varepsilon_{-1}^{(h)}, \ldots),$$

and thus the random variables X_0 and $X_h^{(h)}$ are independent. Stationarity of $\{X_k\}$, independence of X_0 and $X_h^{(h)}$ and the Cauchy-Schwarz inequality yield that

$$\begin{aligned} N E\|\hat{\mu}_N - \mu\|^2 &= \sum_{h=-(N-1)}^{N-1} \left(1 - \frac{|h|}{N}\right) E\langle X_0 - \mu, X_h - \mu\rangle \\ &\le \sum_{h\in\mathbb{Z}} |E\langle X_0 - \mu, X_h - \mu\rangle| \\ &\le E\|X_0 - \mu\|^2 + 2\sum_{h\ge 1} |E\langle X_0 - \mu, X_h - X_h^{(h)}\rangle| \\ &\le \nu_2(X_0 - \mu) \times \left(\nu_2(X_0 - \mu) + 2\sum_{h\ge 1} \nu_2\left(X_h - X_h^{(h)}\right)\right) < \infty. \end{aligned}$$

□

THEOREM 6. *Suppose $\{X_n\} \in L_H^4$ is an L^4–m–approximable sequence with covariance operator C. Then there is some constant $U_X < \infty$, which does not depend on N, such that*

$$E\|\hat{C} - C\|_{\mathcal{S}}^2 \le U_X\, N^{-1}. \tag{25}$$

Before we give the proof, we state the following important result that follows immediately from Theorem 6 and from Lemmas 1 and 2.

THEOREM 7. *Suppose $\{X_n, n \in \mathbb{Z}\} \in L_H^4$ is an L^4–m–approximable sequence and assumption (10) holds. Then, for $1 \le j \le d$,*

$$E\left[|\lambda_j - \hat{\lambda}_j|^2\right] = O(N^{-1}) \quad \text{and} \quad E\left[\|\hat{c}_j\hat{v}_j - v_j\|^2\right] = O(N^{-1}). \tag{26}$$

Theorems 5–7 show that the standard estimates for the functional mean and the FPCs employed for i.i.d data are robust to a sufficiently weak violation of the independence assumption.

PROOF OF THEOREM 6: We assume for simplicity that $EX = 0$. For $k \in \mathbb{Z}$, define the operators $B_k(y) = \langle X_k, y\rangle\, X_k - C(y)$, $y \in H$. Then because B_k are stationary, we have

$$\begin{aligned} E\|\hat{C}_N - C\|_{\mathcal{S}}^2 &= E\left\|\frac{1}{N}\sum_{k=1}^{N} B_k\right\|_{\mathcal{S}}^2 \\ &= \frac{1}{N}\sum_{k=-(N-1)}^{N-1} \left(1 - \frac{|k|}{N}\right) E\langle B_0, B_k\rangle_{\mathcal{S}} \\ &\le \frac{1}{N}\left(E\|B_0\|_{\mathcal{S}}^2 + 2\sum_{k\ge 1} |E\langle B_0, B_k\rangle_{\mathcal{S}}|\right), \end{aligned}$$

and it remains to show that $|E\langle B_0, B_k\rangle_{\mathcal{S}}|$ decays sufficiently fast. We let $\lambda_1 \geq \lambda_2 \geq \cdots$ be the eigenvalues of the operator C and we let $\{e_i\}$ be the corresponding eigenfunctions. It can be readily verified that

$$E\,\langle B_0, B_k\rangle_{\mathcal{S}} = E\,\langle X_0, X_k\rangle^2 - \sum_{j\geq 1}\lambda_j^2, \quad k \geq 1.$$

Furthermore, by using the independence of X_0 and $X_k^{(k)}$, we have

$$E\left\langle X_0, X_k^{(k)}\right\rangle^2 = \sum_{j\geq 1}\lambda_j^2, \quad k \geq 1,$$

showing that

$$E\,\langle B_0, B_k\rangle_{\mathcal{S}} = E\,\langle X_0, X_k\rangle^2 - E\left\langle X_0, X_k^{(k)}\right\rangle^2. \tag{27}$$

For ease of notation, we set $X_k' = X_k^{(k)}$. Then we have

$$\begin{aligned}\left\langle X_0, X_k - X_k'\right\rangle^2 &= \langle X_0, X_k\rangle^2 + \left\langle X_0, X_k'\right\rangle^2 - 2\,\langle X_0, X_k\rangle\left\langle X_0, X_k'\right\rangle \\ &= \langle X_0, X_k\rangle^2 - \left\langle X_0, X_k'\right\rangle^2 - 2\left\langle X_0, X_k - X_k'\right\rangle\left\langle X_0, X_k'\right\rangle.\end{aligned}$$

Thus,

$$\langle X_0, X_k\rangle^2 - \left\langle X_0, X_k'\right\rangle^2 = \left\langle X_0, X_k - X_k'\right\rangle^2 + 2\left\langle X_0, X_k - X_k'\right\rangle\left\langle X_0, X_k'\right\rangle$$

and by repeated application of Cauchy–Schwarz, it follows that

$$\left|E\,\langle X_0, X_k\rangle^2 - E\left\langle X_0, X_k'\right\rangle^2\right| \leq \nu_4^2(X_0)\nu_4^2\left(X_k - X_k'\right) + 2\nu_4^2(X_0)\nu_2(X_0)\nu_2\left(X_k - X_k'\right). \tag{28}$$

Combining (27) and (28) and using the definition of L^4–m–approximability yields the proof of our theorem, with U_X equal to the sum over $k \geq 1$ of the right-hand side of (28). □

4.3. Estimation of the long-run variance

The main results of this section are Corollary 1 and Proposition 1, which state that the long-run variance matrix obtained by projecting the data on the functional principal components can be consistently estimated. We start with some preliminaries, which lead to the main results. For illustration, we present the proof of Lemma 4. More details can be found in the study by Hörmann and Kokoszka (2010).

Let $\{X_n\}$ be a scalar (weakly) stationary sequence. Its long-run variance is defined as $\sigma^2 = \sum_{j\in\mathbb{Z}}\gamma_j$, where $\gamma_j = \mathrm{Cov}(X_0, X_j)$, provided this series is absolutely convergent. Our first Lemma shows that this is the case for L^2–m–approximable sequences.

LEMMA 4. *Suppose* $\{X_n\}$ *is a scalar* L^2*–m–approximable sequence. Then its autocovariance function* $\gamma_j = \mathrm{Cov}(X_0, X_j)$ *is absolutely summable, i.e.,* $\sum_{j=-\infty}^{\infty} |\gamma_j| < \infty$.

PROOF. As we have noted in the proof of Theorem 5, X_0 and $X_j^{(j)}$ are independent, and thus $\mathrm{Cov}(X_0, X_j^{(j)}) = 0$. It follows that $|\gamma_j| \le [EX_0^2]^{1/2}[E(X_j - X_j^{(j)})^2]^{1/2}$, which proves the Lemma. □

The summability of the autocovariances is the fundamental property of weak dependence because $N\mathrm{Var}[\bar{X}_N] \to \sum_{j=-\infty}^{\infty} \gamma_j$, i.e., the variance of the sample mean converges to zero at the rate N^{-1}, the same as for i.i.d. observations. A popular approach to the estimation of the long-run variance is to use the kernel estimator

$$\hat{\sigma}^2 = \sum_{|j|\le q} \omega_q(j)\hat{\gamma}_j, \quad \hat{\gamma}_j = \frac{1}{N}\sum_{i=1}^{N-|j|} (X_i - \bar{X}_N)(X_{i+|j|} - \bar{X}_N).$$

Various weights $\omega_q(j)$ have been proposed and their optimality properties studied, see Andrews (1991) and Anderson (1994), among others. In theoretical work, it is typically assumed that the bandwith q is a deterministic function of the sample size, such that $q = q(N) \to \infty$ and $q = o(N^r)$.

We consider the vector case in which the data are of the form

$$\mathbf{X}_n = [X_{1n}, X_{2n}, \ldots, X_{dn}]^T, \quad n = 1, 2, \ldots, N.$$

The estimation of the mean by the sample mean does not affect the limit of the kernel long-run variance estimators, so *we assume that* $EX_{in} = 0$ and define the autocovariances as

$$\gamma_r(i, j) = E[X_{i0}X_{jr}], \quad 1 \le i, j \le d.$$

If $r \ge 0$, $\gamma_r(i, j)$ is estimated by $N^{-1}\sum_{n=1}^{N-r} X_{in}X_{j,n+r}$, but if $r < 0$, it is estimated by $N^{-1}\sum_{n=1}^{N-|r|} X_{i,n+|r|}X_{j,n}$. The autocovariance matrices are thus

$$\hat{\mathbf{\Gamma}}_r = \begin{cases} N^{-1}\sum_{n=1}^{N-r} \mathbf{X}_n\mathbf{X}_{n+r}^T & \text{if} \quad r \ge 0, \\ N^{-1}\sum_{n=1}^{N-|r|} \mathbf{X}_{n+|r|}\mathbf{X}_n^T & \text{if} \quad r < 0. \end{cases}$$

The variance $\mathrm{Var}[N^{-1}\bar{\mathbf{X}}_N]$ has (i, j)-entry

$$N^{-2}\sum_{m,n=1}^{N} E[X_{im}X_{jn}] = N^{-1}\sum_{|r|<N} \left(1 - \frac{|r|}{N}\right)\gamma_r(i, j),$$

so the long-run variance is

$$\mathbf{\Sigma} = \sum_{r=-\infty}^{\infty} \mathbf{\Gamma}_r, \quad \mathbf{\Gamma}_r := [\gamma_r(i, j),\ 1 \le i, j \le d],$$

and its kernel estimator is

$$\hat{\mathbf{\Sigma}} = \sum_{|r|\le q} \omega_q(r)\hat{\mathbf{\Gamma}}_r. \tag{29}$$

We consider the weights $\omega_q(j) = K(j/q)$, where K is a kernel satisfying the following assumption.

ASSUMPTION 1.

(i) K(0)=1;
(ii) K is a symmetric, Lipschitz function;
(iii) K has a bounded support;
(iv) $\hat{K}$, the Fourier transform of K, is also Lipschitz and integrable.

The following theorem is proven in Horváth and Kokoszka (2012).

THEOREM 8. *Suppose* $\{\mathbf{X}_n\}$ *is an* L^2*–m–approximable sequence. If Assumption 1 holds and* $q \to \infty$, $q/N \to 0$, *then* $\hat{\mathbf{\Sigma}}_N \xrightarrow{P} \mathbf{\Sigma}$.

In contrast to many classical results, see e.g., Newey and West (1987), Theorem 8 does not impose fourth-order conditions and replaces mixing or linearity conditions by L^2*–m–*approximability. Assumption 1 is standard, except its condition (iv). The following example shows that it holds for the Bartlett kernel; the Fourier transforms of other commonly used kernels are smoother and decay faster.

Example 3. The Bartlett kernel is

$$K(s) = \begin{cases} 1 - |s|, & |s| \le 1, \\ 0, & \text{otherwise} \end{cases}$$

This kernel clearly satisfies parts (i)–(iii) of Assumption 1. Its Fourier transform is

$$\hat{H}(u) = \left\{ \frac{1}{\pi u} \sin\left(\frac{u}{2}\right) \right\}^2.$$

Thus, to verify part (iv), we must check that the function

$$F(t) = \left\{ \frac{\sin(t)}{t} \right\}^2$$

is integrable and Lipschitz. The integrability follows because $|F(t)| \le t^{-2}$ and $F(t) \to 1$, as $t \to 0$.

The derivative of F for $t \neq 0$ is

$$F'(t) = \frac{2\sin(t)}{t}\left\{\frac{t\cos(t) - \sin(t)}{t^2}\right\}.$$

This function is clearly bounded outside any neighborhood of zero. Using the Taylor expansion of the sine and cosine functions, it is easy to verify that $F'(t) = o(t)$, as $t \to 0$. In a similar fashion, one can verify that $F(t) - F(0) = o(t^2)$, as $t \to 0$. Thus, F is Lipschitz on the whole line.

We are now able to turn to functional data. Suppose $\{X_n\} \in L_H^2$ is a zero mean sequence and $e_1, e_2, \ldots, e_d$ is any set of orthonormal functions in H. Define $X_{in} = \langle X_n, e_i\rangle$, $\mathbf{X}_n = [X_{1n}, X_{2n}, \ldots, X_{dn}]^T$ and $\Gamma_r = \mathrm{Cov}(\mathbf{X}_0, \mathbf{X}_r)$. A direct verification shows that if $\{X_n\}$ is L^p–m–approximable, then so is the vector sequence $\{\mathbf{X}_n\}$. Thus, we obtain the following corollary.

COROLLARY 1. *(a) If $\{X_n\} \in L_H^2$ is an L^2–m–approximable sequence, then the series $\sum_{r=-\infty}^{\infty} \mathbf{\Gamma}_r$ converges absolutely. (b) If, in addition, Assumption 1 holds and $q \to \infty$ with $q = o(N)$, then $\hat{\mathbf{\Sigma}} \xrightarrow{P} \mathbf{\Sigma}$.*

In Corollary 1, the functions $e_1, e_2, \ldots, e_d$ form an arbitrary orthonormal deterministic basis. In many applications, a random basis consisting of the EFPC's $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_d$ is used. The scores with respect to this basis are defined by

$$\hat{\eta}_{\ell i} = \left\langle X_i - \bar{X}_N, \hat{v}_\ell \right\rangle, \quad 1 \leq \ell \leq d.$$

To use the results established so far, it is convenient to decompose the stationary sequence $\{X_n\}$ into its mean and a zero mean process, i.e., we set $X_n = \mu + Y_n$, where $EY_n = 0$. We introduce the unobservable quantities

$$\beta_{\ell n} = \langle Y_n, v_\ell\rangle, \quad \hat{\beta}_{\ell n} = \left\langle Y_n, \hat{v}_\ell \right\rangle, \quad 1 \leq \ell \leq d.$$

The following proposition is useful in the development of asymptotic arguments for many statistical procedures for functional time series. The boldface symbols refer to the vectors with the coordinates just defined, and $\hat{\mathbf{\Sigma}}(\boldsymbol{\delta})$ is the estimator (29) calculated from the observation vectors $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_N$.

PROPOSITION 1. *Let $\hat{\mathbf{C}} = \mathrm{diag}(\hat{c}_1, \ldots, \hat{c}_d)$, with $\hat{c}_i = \mathrm{sign}(\langle v_i, \hat{v}_i\rangle)$. Suppose $\{X_n\} \in L_H^4$ is L^4–m–approximable and that (10) holds. Assume further that Assumption 1 holds with a stronger condition $q^4/N \to 0$. Then*

$$|\hat{\mathbf{\Sigma}}(\boldsymbol{\beta}) - \hat{\mathbf{\Sigma}}(\hat{\mathbf{C}}\hat{\boldsymbol{\beta}})| = o_P(1) \quad \textit{and} \quad |\hat{\mathbf{\Sigma}}(\hat{\boldsymbol{\eta}}) - \hat{\mathbf{\Sigma}}(\hat{\boldsymbol{\beta}})| = o_P(1). \tag{30}$$

The point of Proposition 1 is that if $\hat{\mathbf{\Sigma}}(\boldsymbol{\beta})$ is consistent under some conditions, e.g., those stated in Theorem 8, then so is $\hat{\mathbf{\Sigma}}(\hat{\boldsymbol{\eta}})$. Before presenting the proof of Proposition 1, we note that for functional data, it is also often useful to consider

the long-run covariance kernel, which is defined in terms of the kernels $c_h(t,s) = \mathrm{Cov}(X_i(t), X_{i+h}(t))$. Kernel estimators can be defined analogously to (29), and their consistency can be established under L^2–m–approximability and additional technical conditions; we refer to Horváth et al. (2012).

Proof of Proposition 1. We only prove the left relation in (30). We introduce the constant

$$\kappa := \sup_{q\geq 1} \frac{1}{q} \sum_{j=-q}^{q} w_q(j),$$

which by Assumption 1 is finite (and converges to $2\int_{-1}^{1} K(x)dx$). The element in the k-th row and ℓ-th column of $\hat{\mathbf{\Sigma}}(\boldsymbol{\beta}) - \hat{\mathbf{\Sigma}}(\hat{\mathbf{C}}\hat{\boldsymbol{\beta}})$ is given by

$$\sum_{h=0}^{q} \frac{w_q(h)}{N} \sum_{1\leq n\leq N-h} \left(\beta_{kn}\beta_{\ell,n+h} - \hat{c}_k\hat{\beta}_{kn}\hat{c}_\ell\hat{\beta}_{\ell,n+h}\right)$$
$$+ \sum_{h=1}^{q} \frac{w_q(h)}{N} \sum_{1\leq n\leq N-h} \left(\beta_{k,n+h}\beta_{\ell,n} - \hat{c}_k\hat{\beta}_{k,n+h}\hat{c}_\ell\hat{\beta}_{\ell,n}\right). \tag{31}$$

For reasons of symmetry, it suffices to study (31), which can be decomposed into

$$\sum_{h=0}^{q} \frac{w_q(h)}{N} \sum_{1\leq n\leq N-h} \beta_{kn}\left(\beta_{\ell,n+h} - \hat{c}_\ell\hat{\beta}_{\ell,n+h}\right)$$
$$+ \sum_{h=0}^{q} \frac{w_q(h)}{N} \sum_{1\leq n\leq N-h} \hat{c}_\ell\hat{\beta}_{\ell,n+h}\left(\beta_{kn} - \hat{c}_k\hat{\beta}_{kn}\right). \tag{32}$$

As both summands above can be treated similarly, we will only treat (32). For any $\varepsilon > 0$, we have

$$P\left(\left|\sum_{h=0}^{q} \frac{w_q(h)}{N} \sum_{1\leq n\leq N-h} \beta_{kn}\left(\beta_{\ell,n+h} - \hat{c}_\ell\hat{\beta}_{\ell,n+h}\right)\right| > \varepsilon\kappa\right)$$
$$\leq P\left(\left|\sum_{h=0}^{q} \frac{w_q(h)}{N} \sum_{1\leq n\leq N-h} \beta_{kn}\left(\beta_{\ell,n+h} - \hat{c}_\ell\hat{\beta}_{\ell,n+h}\right)\right| > \frac{\varepsilon}{q}\sum_{h=0}^{q} w_q(h)\right)$$
$$\leq \sum_{h=0}^{q} P\left(\frac{1}{N}\left|\sum_{1\leq n\leq N-h} \beta_{kn}\left(\beta_{\ell,n+h} - \hat{c}_\ell\hat{\beta}_{\ell,n+h}\right)\right| > \frac{\varepsilon}{q}\right). \tag{33}$$

To show that (33) tends to 0 as $N \to \infty$, we introduce a slowly increasing sequence $\alpha_N \to \infty$, such that $q^4\alpha_N/N \to 0$, and we let C_0, such that $N \max_{1\le\ell\le d} E\|v_\ell - \hat{c}_\ell\hat{v}_\ell\|^2 \le C_0$. By Cauchy–Schwarz and Markov inequality, we have

$$
\begin{aligned}
&P\left(\left|\sum_{1\le n\le N-h}\beta_{kn}\left(\beta_{\ell,n+h}-\hat{c}_\ell\hat{\beta}_{\ell,n+h}\right)\right| > \frac{\varepsilon N}{q}\right)\\
&\le P\left(\sum_{n=1}^N \beta_{kn}^2 \sum_{n=1}^N\left(\beta_{\ell n}-\hat{c}_\ell\hat{\beta}_{\ell n}\right)^2 > \frac{\varepsilon^2 N^2}{q^2}\right)\\
&\le P\left(\frac{1}{N}\sum_{n=1}^N \beta_{kn}^2 > q\alpha_N\right) + P\left(\frac{1}{N}\sum_{n=1}^N\left(\beta_{\ell n}-\hat{c}_\ell\hat{\beta}_{\ell n}\right)^2 > \frac{\varepsilon^2}{q^3\alpha_N}\right)\\
&\le \frac{E\beta_{k1}^2}{q\alpha_N} + P\left(\frac{1}{N}\sum_{n=1}^N \|Y_n\|^2\|v_\ell-\hat{c}_\ell\hat{v}_\ell\|^2 > \frac{\varepsilon^2}{q^3\alpha_N}\right)\\
&\le \frac{E\|Y_1\|^2}{q\alpha_N} + P\left(\frac{1}{N}\sum_{n=1}^N \|Y_n\|^2 > 2E\|Y_1\|^2\right)\\
&\quad + P\left(\|v_\ell-\hat{c}_\ell\hat{v}_\ell\|^2 > \frac{\varepsilon^2}{2E\|Y_1\|^2 q^3\alpha_N}\right)\\
&\le \frac{E\|Y_1\|^2}{q\alpha_N} + \frac{\mathrm{Var}\left(\frac{1}{N}\sum_{n=1}^N\|Y_n\|^2\right)}{E^2\|Y_1\|^2} + \frac{2C_0\,E\|Y_1\|^2 q^3\alpha_N}{N\varepsilon^2}.
\end{aligned}
$$

It can be easily shown that for U, V in L_H^4

$$
\nu_2\left(\|U\|^2 - \|V\|^2\right) \le \nu_4^2(U-V) + 2\left\{\nu_4(U)+\nu_4(V)\right\}\nu_4(U-V).
$$

An immediate consequence is that L^4–m–approximability of $\{Y_n\}$ implies L^2–m–approximability of the scalar sequence $\{\|Y_n\|^2\}$. A basic result for stationary sequences gives

$$
\mathrm{Var}\left(\frac{1}{N}\sum_{n=1}^N\|Y_n\|^2\right) \le \frac{1}{N}\sum_{h\in\mathbb{Z}}\left|\mathrm{Cov}\left(\|Y_0\|^2, \|Y_h\|^2\right)\right|,
$$

where by Lemma 4, the autocovariances are absolutely summable. Hence, the summands in (33) are bounded by

$$
C_1\left\{\frac{1}{q\alpha_N} + \frac{1}{N} + \frac{q^3\alpha_N}{N\varepsilon^2}\right\},
$$

where the constant C_1 depends only on the law of $\{Y_n\}$. The proof of the proposition follows immediately from our assumptions on q and α_N.

5. Further reading

All topics discussed in this survey are presented in detail in Horváth and Kokoszka (2012). Bosq (2000) contains theoretical foundations for most results of Sections 2 and 3. Ramsay and Silverman (2005) provide an introduction to many fundamental concepts of FDA, while Ramsay et al. (2009) focus on implementation in `R` and `MATLAB`.

A topic of particular importance in time series analysis is *change point detection*. Most approaches to modeling time series assume that the data follow one model. If the stochastic structure of the data changes at some time point(s), both exploratory and inferential tools produce misleading results. Berkes et al. (2009) study the problem of testing for a change in the mean function assuming that the curves are collected over time, but are independent. Hörmann and Kokoszka (2010) extend their procedure to L^4–m–approximable functional time series. Asymptotic distributions of related change point estimators are studied in Aue et al. (2009). Horváth et al. (2009) develop a test for a change point in the autoregressive operator Ψ in the FAR(1) model. Gabrys et al. (2010a) use the framework of FDA to detect changes in the intraday volatility pattern, while Aston and Kirch (2011a,b) consider fMRI data.

A central topic in FDA is the *functional linear model* of the for $Y_n = \Psi(X_n) + \varepsilon_n$. In its most general form, the responses Y_n, the regressors X_n and the errors ε_n are functions, and Ψ is an integral kernel operator. Very extensive research is available under the assumption that the cases (X_n, Y_n) are independent and the errors ε_n are independent. Hörmann and Kokoszka (2010) showed that an estimator for Ψ developed under i.i.d. assumption remains consistent if the X_n form an L^4–m–approximable time series. Gabrys et al. (2010b) developed procedures to test the assumption of i.i.d. ε_n against the alternative that the ε_n are correlated. Gabrys and Kokoszka (2007) developed a similar test, which is, however, applicable to directly observable curves, not to unobservable errors.

Dependence between curves plays a central role also for *spatial* functional data. In this context, we observe curves at many spatial locations, for example, the precipitation over many decades at a number of measurement stations. In addition to the dependence between curves, spatial distribution of the locations must also be taken into account to develop informative statistical procedures. Hörmann and Kokoszka (2012) develop asymptotic theory for the estimation of the mean function and the FPCs for such data. Gromenko et al. (2011) propose and compare several estimation procedures, which improve on the standard simple mean and the EFPC's defined in Section 2. The focus of research for spatially indexed curves has, however, been kriging (spatial prediction), see Delicado et al. (2010), Nerini et al. (2010), Giraldo et al. (2010), and Bel et al. (2011).

Acknowledgments

This research was partially supported by the National Science Foundation, the Banque nationale de Belgique, and Communauté française de Belgique – Actions de Recherche Concertées (2010–2015).

References

Akhiezier, N.I., Glazman, I.M., 1993. Theory of Linear Operators in Hilbert Space. Dover, New York.

Anderson, T.W., 1994. The Statistical Analysis of Time Series. Wiley & Sons.

Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–58.

Antoniadis, A., Paparoditis, E., Sapatinas, T., 2006. A functional wavelet–kernel approach for time series prediction. J. Roy. Stat. Soc. B 68, 837–57.

Aston, J.A.D., Kirch, C., 2011a. Detecting and estimating epidemic changes in dependent functional data. CRiSM Research Report 11–07. University of Warwick.

Aston, J.A.D., Kirch, C., 2011b. Estimation of the distribution of change-points with application to fMRI data. CRiSM Research Reports. University of Warwick.

Aue, A., Gabrys, R., Horváth, L., Kokoszka, P., 2009. Estimation of a change–point in the mean function of functional data. J. Multivariate Anal. 100, 2254–69.

Aue, A., Hörmann, S., Horváth, L., Reimherr, M., 2009. Break detection in the covariance structure of multivariate time series models. Ann. Stat. 37, 4046–87.

Bel, L., Bar-Hen, A., Cheddadi, R., Petit, R., 2011. Spatio-temporal functional regression on paleo–ecological data. J. Appl. Stat. 38, 695–704.

Berkes, I., Gabrys, R., Horváth, L., Kokoszka, P., 2009. Detecting changes in the mean of functional observations. J. Roy. Stat. Soc. B 71, 927–46.

Berkes, I., Hörmann, S., Schauer, J., 2011. Split invariance principles for stationary processes. Ann. Probab. 39, 2441–2473.

Besse, P., Cardot, H., Stephenson, D., 2000. Autoregressive forecasting of some functional climatic variations. Scand. J. Stat. 27, 673–87.

Billingsley, P., 1968. Convergence of Probability Measures. Wiley, New York.

Bosq, D., 2000. Linear Processes in Function Spaces. Springer, New York.

Bradley, R.C., 2007. Introduction to Strong Mixing Conditions, vol. 1,2,3. Kendrick Press.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods. Springer, New York.

Dauxois, J., Pousse, A., Romain, Y., 1982. Asymptotic theory for principal component analysis of a vector random function. J. Multivariate Anal. 12, 136–54.

Debnath, L., Mikusinski, P., 2005. Introduction to Hilbert Spaces with Applications. Elsevier.

Delicado, P., Giraldo, R., Comas, C., Mateu, J., 2010. Statistics for spatial functional data: some recent contributions. Environmetrics 21, 224–39.

Diaconis, P., Freeman, D., 1999. Iterated random functions. SIAM Rev. 41, 45–76.

Didericksen, D., Kokoszka, P., Zhang, X., in press. Empirical properties of forecasts with the functional autoregressive model. Comput. Stat.

Doukhan, P., 1994. Mixing: Properties and Examples. Lecture Notes in Statistics. Springer.

Doukhan, P., Louhichi, S., 1999. A new weak dependence and applications to moment inequalities. Stoch. Processes Appl. 84, 313–43.

Gabrys, R., Hörmann, S., Kokoszka, P., 2010a. Monitoring the intraday volatility pattern. Technical Report. Utah State University.

Gabrys, R., Horváth, L., Kokoszka, P., 2010b. Tests for error correlation in the functional linear model. J. Am. Stat. Assoc. 105, 1113–25.

Gabrys, R., Kokoszka, P., 2007. Portmanteau test of independence for functional observations. J. Am. Stat. Assoc. 102, 1338–48.

Giraldo, R., Delicado, P., Mateu, J., 2010. Geostatistics for functional data: An ordinary kriging approach. Environ. Ecol. Stat. 18, 411–426.

Gohberg, I., Golberg, S., Kaashoek, M.A., 1990. Classes of Linear Operators. Operator Theory: Advances and Applications, vol. 49. Birkhaüser.

Gromenko, O., Kokoszka, P., Zhu, L., Sojka, J., 2011. Estimation and testing for spatially distributed curves with application to ionospheric and magnetic field trends. Technical Report. Utah State University.

Hörmann, S., 2008. Augmented GARCH sequences: Dependence structure and asymp totics. Bernoulli 14, 543–61.

Hörmann, S., Kokoszka, P., 2010. Weakly dependent functional data. Ann. Stat., 38, 1845–84.

Hörmann, S., Kokoszka, P., 2012. Consistency of the mean and the principal components of spatially distributed functional data. Bernoulli.

Horváth, L., Hušková, M., Kokoszka, P., 2009. Testing the stability of the functional autoregressive process. J. Multivariate Anal., 101, 352–67.

Horváth, L., Kokoszka, P., 2012. Inference for Functional Data with Applications. Springer Series in Statistics. Springer.

Horváth, L., Kokoszka, P., Reeder, R., 2012. Estimation of the mean of functional time series and a two sample problem. J. Roy. Stat. Soc. B.

Johnstone, I.M., Lu, A.Y., 2009. On consistency and sparcity for principal components analysis in high dimensions. J. Am. Stat. Assoc. 104, 682–93.

Kargin, V., Onatski, A., 2008. Curve forecasting by functional autoregression. J. Multivariate Anal. 99, 2508–26.

Kokoszka, P., Zhang, X., 2010. Improved estimation of the kernel of the functional autoregressive process. Technical Report. Utah State University.

Nerini, D., Monestiez, P., Mantéa, C., 2010. Cokriging for spatial functional data. J. Multivariate Anal. 101, 409–18.

Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–8.

Pötscher, B., Prucha, I., 1997. Dynamic Non–linear Econonometric Models. Asymptotic Theory. Springer.

Ramsay, J., Hooker, G., Graves, S., 2009. Functional Data Analysis with R and MATLAB. Springer.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. Springer.

Riesz, F., Sz.-Nagy, B., 1990. Functional Analysis. Dover.

Wu, W., 2005. Nonlinear system theory: another look at dependence. Proc. Natl. Acad. Sci. USA 102, 14150–4.

Wu, W., 2007. Strong invariance principles for dependent random variables. Ann. Probab 35, 2294–320.

Wu, W.B., Shao, X., 2004. Limit theorems for iterated random functions. J. Appl. Probab. 41, 425–36.

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00008-9

Covariance Matrix Estimation in Time Series

Wei Biao Wu and Han Xiao
Department of Statistics, The University of Chicago, Chicago, IL 60637, USA

Abstract

Covariances play a fundamental role in the theory of time series, and they are critical quantities that are needed in both spectral and time domain analysis. Estimation of covariance matrices is needed in the construction of confidence regions for unknown parameters, hypothesis testing, principal component analysis, prediction, discriminant analysis, among others. In this chapter, we consider both low- and high-dimensional covariance matrix estimation problems and present a review for asymptotic properties of sample covariances and covariance matrix estimates. In particular, we shall provide an asymptotic theory for estimates of high-dimensional covariance matrices in time series and a consistency result for covariance matrix estimates for estimated parameters.

Keywords: high-dimensional inference, stationary process, spectral density estimation, Heteroscedasticity and Autocorrelation Consistent, regularization.

1. Introduction

Covariances and covariance matrices play a fundamental role in the theory and practice of time series. They are critical quantities that are needed in both spectral and time domain analysis. One encounters the issue of covariance matrix estimation in many problems, for example, the construction of confidence regions for unknown parameters, hypothesis testing, principal component analysis, prediction, discriminant analysis, among others. It is particularly relevant in time series analysis in which the observations are dependent, and the covariance matrix characterizes the second-order dependence of the process. If the underlying process is Gaussian, then the covariances

completely capture its dependence structure. In this chapter, we shall provide an asymptotic distributional theory for sample covariances and convergence rates for covariance matrix estimates of time series.

In Section 2 , we shall present a review for asymptotic theory for sample covariances of stationary processes. In particular, the limiting behavior of sample covariances at both small and large lags is discussed. The obtained result is useful for constructing consistent covariance matrix estimates for stationary processes. We shall also present a uniform convergence result so that one can construct simultaneous confidence intervals for covariances and perform tests for white noises. In that section, we also introduce dependence measures that are necessary for asymptotic theory for sample covariances.

Sections 3 and 4 concern estimation of covariance matrices, the main theme of the paper. There are basically two types of covariance matrix estimation problems: the first one is the estimation of covariance matrices of some estimated finite-dimensional parameters. For example, given a sequence of observations $Y_1, \ldots, Y_n$, let $\hat{\theta}_n = \hat{\theta}_n(Y_1, \ldots, Y_n)$ be an estimate of the unknown parameter vector $\theta_0 \in \mathbb{R}^d$, $d \in \mathbb{N}$, which is associated with the process (Y_i). For statistical inference of θ_0, one would like to estimate the $d \times d$ covariance matrix $\Sigma_n = \mathrm{cov}(\hat{\theta}_n)$. For example, with an estimate of Σ_n, confidence regions for θ_0 can be constructed and hypotheses regarding θ_0 can be tested. We generically call such problems as low-dimensional covariance matrix estimation problem since the dimension d is assumed to be fixed and it does not grow with n.

For the second type, let $(X_1, \ldots, X_p)$ be a p-dimensional random vector with $E(X_i^2) < \infty$, $i = 1, \ldots, p$; let $\gamma_{i,j} = \mathrm{cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$, $1 \le i, j \le p$, be its covariance function. The problem is to estimate the $p \times p$ dimensional matrix

$$\Sigma_p = (\gamma_{i,j})_{1 \le i,j \le p}. \tag{1}$$

A distinguished feature of such type of problem is that the dimension p can be very large. Techniques and asymptotic theory for high-dimensional covariance matrix estimates are quite different from the low-dimensional ones. On the other hand, however, we can build the asymptotic theory for both cases based on the same framework of causal processes and the physical dependence measure proposed in the study by Wu (2005).

The problem of low-dimensional covariance matrix estimation is discussed in Section 3. In particular, we consider the latter problem in the context of sample means of random vectors and estimates of linear regression parameters. We shall review the classical theory of Heteroscedasticity and Autocorrelation Consistent (HAC) covariance matrix estimates of White (1980), Newey and West (1987), Andrews (1991), Andrews and Monahan (1992), de Jong and Davidson (2000), and among others. In comparison with those traditional result, an interesting feature of our asymptotic theory is that we impose very mild moment conditions. Additionally, we do not need the strong mixing conditions and the cumulant summability conditions that are widely used in the literature (Andrews, 1991; Rosenblatt, 1985). For example, for consistency of covariance matrix estimates, we only require the existence of 2 or $(2 + \epsilon)$ moments, where $\epsilon > 0$ can be very small, while in the classical theory one typically needs the existence of 4 moments. The imposed dependence conditions are easily verifiable and

they are optimal in certain sense. In the study of the convergence rates of the estimated covariance matrices, since the dimension is finite, all commonly used norms (e.g., the operator norm, the Frobenius norm, and the $\mathcal{L}^1$ norm) are equivalent and the convergence rates do not depend on the norm that one chooses.

Section 4 deals with the second-type covariance matrix estimation problem in which p can be big. Due to the high dimensionality, the norms mentioned above are no longer equivalent. Additionally, unlike the lower dimensional case, the sample covariance matrix estimate is no longer consistent. Hence suitable regularization procedures are needed so that the consistency can be achieved. In Section 4, we shall use the operator norm: for an $p \times p$ matrix A, let

$$\rho(A) = \sup_{v:|v|=1} |Av| \tag{2}$$

be the operator norm (or spectral radius), where for a vector $v = (v_1, \ldots, v_p)^\top$, its length $|v| = (\sum_{i=1}^p v_i^2)^{1/2}$. Section 4 provides an exact order of the operator norm of the sample autocovariance matrix and the convergence rates of regularized covariance matrix estimates. We shall review the regularized covariance matrix estimation theory of Bickel and Levina (2008a,b), the Cholesky decomposition theory in Pourahmadi (1999), Wu and Pourahmadi (2003), and among others, and the parametric covariance matrix estimation using generalized linear models. Suppose one has n independent and identically distributed (i.i.d.) realizations of $(X_1, \ldots, X_p)$. In many situations, p can be much larger than n, which is the so-called large p small n problem. Bickel and Levina (2008a) showed that the banded covariance matrix estimate is consistent in operator norm if X_i's have a very short tail and the growth speed of the number of replicates n can be such that $\log(p) = o(n)$. In many time series applications, however, there is only one realization available, namely $n = 1$. In Section 4, we shall consider high-dimensional matrix estimation for both one and multiple realizations. In the former case, we assume stationarity and use sample autocovariance matrix. A banded version of the sample autocovariance matrix can be consistent.

2. Asymptotics of sample covariances

In this section, we shall introduce the framework of stationary causal process, its associated dependence measures, and an asymptotic theory for sample covariances. If the process (X_i) is stationary, then $\gamma_{i,j}$ can be written as $\gamma_{i-j} = \operatorname{cov}(X_0, X_{i-j})$, and $\Sigma_n = (\gamma_{i-j})_{1\le i,j\le n}$ is then a Toeplitz matrix. Assuming at the outset that $\mu = EX_i = 0$. To estimate Σ_n, it is natural to replace γ_k in Σ_n by the sample version

$$\hat{\gamma}_k = \frac{1}{n} \sum_{i=1+|k|}^{n} X_i X_{i-|k|}, \quad 1 - n \le k \le n - 1. \tag{3}$$

If $\mu = EX_i$ is not known, then we can modify (3) by

$$\tilde{\gamma}_k = \frac{1}{n} \sum_{i=1+|k|}^{n} (X_i - \bar{X}_n)(X_{i-|k|} - \bar{X}_n), \quad \text{where } \bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}. \tag{4}$$

Section 4.4 concerns estimation of Σ_n, and asymptotic properties of $\tilde{\gamma}_k$ will be useful for deriving convergence rates of estimates of Σ_n.

There is a huge literature on asymptotic properties of sample covariances. For linear processes, this problem has been studied in the work of Priestley (1981), Brockwell and Davis (1991), Hannan (1970, 1976), Anderson (1971), Hall and Heyde (1980), Hosking (1996), Phillips and Solo (1992), Wu and Min (2005), and Wu et al. (2010). If the lag k is fixed and bounded, then $\hat{\gamma}_k$ is basically the sample average of the stationary process of lagged products $(X_i X_{i-|k|})$ and one can apply the limit theory for strong mixing processes; see the study by Ibragimov and Linnik (1971), Eberlein and Taqqu (1986), Doukhan (1994), and Bradley (2007).

The asymptotic problem for $\hat{\gamma}_k$ with unbounded k is important since, with that, one can assess the dependence structure of the underlying process by examining its autocovariance function (ACF) plot at large lags. For example, if the time series is a moving average process with an unknown order, then as a common way one can estimate the order by checking its ACF plot. However, the latter problem is quite challenging if the lag k can be unbounded. Keenan (1997) derived a central limit theorem under the very restrictive lag condition $k_n \to \infty$ with $k_n = o(\log n)$ for strong mixing processes whose mixing coefficients decay geometrically fast. A larger range of k_n is allowed in the study by Harris et al. (2003). However, they assume that the process is linear. Wu (2009) dealt with nonlinear processes and the lag condition can be quite weak.

To study properties of sample covariances and covariance matrix estimates, it is necessary to impose appropriate structural conditions on (X_i). Here we assume that it is of the form

$$X_i = H(\varepsilon_i, \varepsilon_{i-1}, \ldots), \tag{5}$$

where ε_j, $j \in \mathbb{Z}$, are i.i.d. and H is a measurable function such that X_i is properly defined. The framework (5) is very general and it includes many widely used linear and nonlinear processes (Wu, 2005). Wiener (1958) claimed that, for every stationary purely nondeterministic process $(X_j)_{j\in\mathbb{Z}}$, there exists i.i.d. uniform$(0, 1)$ random variables ε_j, and a measurable function H such that (5) holds. The latter claim, however, is generally not true; see the work done by Rosenblatt (2009), Ornstein (1973), and Kalikow (1982). Nonetheless, the above construction suggests that the class of processes that (5) represents can be very huge. See the study by Borkar (1993), Tong (1990), Kallianpur (1981), Ornstein (1973), and Rosenblatt (2009) for more historical backgrounds on the above stochastic realization theory. See also the study by Wu (2011) for examples of stationary processes that are of form (5).

Following the study by Priestley (1988) and Wu (2005), we can view (X_i) as a physical system with $(\varepsilon_j, \varepsilon_{j-1}, \ldots)$ (resp. X_i) being the input (resp. output) and H being the transform, filter, or data-generating mechanism. Let the shift process

$$\mathcal{F}_i = (\varepsilon_i, \varepsilon_{i-1}, \ldots). \tag{6}$$

Let $(\varepsilon_i')_{i\in\mathbb{Z}}$ be an i.i.d. copy of $(\varepsilon_i)_{i\in\mathbb{Z}}$. Hence $\varepsilon_i', \varepsilon_j, i, j \in \mathbb{Z}$, are i.i.d. For $l \le j$, define

$$\mathcal{F}_{j,l}^* = (\varepsilon_j, \ldots, \varepsilon_{l+1}, \varepsilon_l', \varepsilon_{l-1}, \ldots).$$

If $l > j$, let $\mathcal{F}_{j,l}^* = \mathcal{F}_j$. Define the projection operator

$$\mathcal{P}_j \cdot = E(\cdot|\mathcal{F}_j) - E(\cdot|\mathcal{F}_{j-1}). \tag{7}$$

For a random variable X, we say $X \in \mathcal{L}^p$ ($p > 0$) if $\|X\|_p := (E|X|^p)^{1/p} < \infty$. Write the $\mathcal{L}^2$ norm $\|X\| = \|X\|_2$. Let $X_i \in \mathcal{L}^p$, $p > 0$. For $j \geq 0$, define the physical (or functional) dependence measure

$$\delta_p(j) = \|X_j - X_j^*\|_p, \quad \text{where} X_j^* = H(\mathcal{F}_{j,0}). \tag{8}$$

Note that X_j^* is a coupled version of X_j with ε_0 in the latter being replaced by ε_0'. The dependence measure (8) greatly facilitates asymptotic study of random processes. In many cases, it is easy to work with and it is directly related to the underlying data-generating mechanism of the process. For $p > 0$, introduce the p-stability condition

$$\Delta_p := \sum_{i=0}^{\infty} \delta_p(i) < \infty. \tag{9}$$

As explained in Wu (2005), (9) means that the cumulative impact of ε_0 on the process $(X_i)_{i\geq 0}$ is finite, thus suggesting short-range dependence. If the above condition is barely violated, then the process (X_i) may be long-range dependent and the spectral density no longer exists. For example, let $X_n = \sum_{j=0}^{\infty} a_j \varepsilon_{n-j}$ with $a_j \sim j^{-\beta}$, $1/2 < \beta$, and ε_i are i.i.d., then $\delta_p(k) = |a_k|\|\varepsilon_0 - \varepsilon_0'\|_p$ and (9) is violated if $\beta < 1$. The latter is a well-known long-range dependent process. If K is a Lipschitz continuous function, then for the process $X_n = K(\sum_{j=0}^{\infty} a_j \varepsilon_{n-j})$, its physical dependence measure $\delta_p(k)$ is also of order $O(|a_k|)$. Wu (2011) also provides examples of Volterra processes, nonlinear AR(p) and AR(∞) processes for which $\delta_p(i)$ can be computed and (9) can be verified.

For a matrix A, we denote its transpose by $A^\top$.

THEOREM 1. *(Wu, 2009, 2011) Let $k \in \mathbb{N}$ be fixed and $E(X_i) = 0$; let $Y_i = (X_i, X_{i-1}, \ldots, X_{i-k})^\top$ and $\Gamma_k = (\gamma_0, \gamma_1, \ldots, \gamma_k)^\top$.*

(i) Assume $X_i \in \mathcal{L}^p$, $2 < p \leq 4$, and (9) holds with this p. Then for all $0 \leq k \leq n-1$, we have

$$\left\|\hat{\gamma}_k - (1 - k/n)\gamma_k\right\|_{p/2} \leq \frac{4n^{2/p-1}\|X_1\|_p \Delta_p}{p-2}. \tag{10}$$

(ii) Assume $X_i \in \mathcal{L}^4$ and (9) holds with $p = 4$. Then as $n \to \infty$

$$\sqrt{n}(\hat{\gamma}_0 - \gamma_0, \hat{\gamma}_1 - \gamma_1, \ldots, \hat{\gamma}_k - \gamma_k) \Rightarrow N[0, E(D_0 D_0^\top)], \tag{11}$$

where $D_0 = \sum_{i=0}^{\infty} \mathcal{P}_0(X_i Y_i) \in \mathcal{L}^2$ and $\mathcal{P}_0$ is the projection operator defined by (7).

(iii) Let $l_n \to \infty$ and assume (9) with $p = 4$. Then we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [X_i Y_{i-l_n} - E(X_{l_n} Y_0)] \Rightarrow N(0, \Sigma_h), \tag{12}$$

where Σ_h is an $h \times h$ matrix with entries

$$\sigma_{ab} = \sum_{j\in\mathbb{Z}} \gamma_{j+a}\gamma_{j+b} = \sum_{j\in\mathbb{Z}} \gamma_j\gamma_{j+b-a} =: \sigma_{0,a-b}, \quad 1 \le a, b \le h, \tag{13}$$

and if additionally $l_n/n \to 0$, then

$$\sqrt{n}[(\hat{\gamma}_{l_n}, \ldots, \hat{\gamma}_{l_n-h+1})^\top - (\gamma_{l_n}, \ldots, \gamma_{l_n-h+1})^\top] \Rightarrow N(0, \Sigma_h). \tag{14}$$

An attractive feature of Theorem 1 is that it provides an explicit error bound (10), which in many cases is sharp up to a multiplicative constant. This is a significant merit for our framework of causal processes with functional or physical dependence measures. See also other theorems in later sections. In (11) and (12), we give explicit forms of the asymptotic covariance matrices, and they can be estimated by using the techniques in Section 3.

Theorem 1 suggests that, at large lags, $\sqrt{n}(\hat{\gamma}_k - E\hat{\gamma}_k)$ behaves asymptotically as $\sum_{j\in\mathbb{Z}} \gamma_j\eta_{k-j}$, where η_j are i.i.d. standard normal random variables. Wu (2011) discussed the connection with Bartlett's (1946) approximate expressions of covariances of estimated covariances. This result implies that the sample covariance $\hat{\gamma}_k$ can be a bad estimate for γ_k if γ_k is small, due to the weak signal-to-noise ratio. Specifically, if k_n is such that $\gamma_{k_n} = o(n^{-1/2})$, then the sample covariance $\hat{\gamma}_{k_n}$ has an asymptotic mean squared error (MSE) σ_{00}/n, which is larger than $\gamma_{k_n}^2$. Note that $\gamma_{k_n}^2$ is the MSE of the trivial estimate $\breve{\gamma}_{k_n} = 0$. The MSE of the truncated estimate of the form $\bar{\gamma}_k = \hat{\gamma}_k \mathbf{1}_{|\hat{\gamma}_k|\ge c_n}$, where $c_n = c/\sqrt{n}$ for some constant $c > 0$, can reach the minimum order of magnitude $O[\min(1/n, r_n^2)]$. Similar truncation ideas are used in the study by Lumley and Heagerty (1999) and Bickel and Levina (2008b). The latter paper deals with thresholded covariance matrix estimators; see Section 4.4.

As a popular way to test the existence of correlations of a process, one checks its ACF plot. Testing of correlations involves testing multiple hypotheses H_0:$\gamma_1 = \gamma_2 = \ldots = 0$. The multiplicity issue should be adjusted if the number of lags is unbounded. To develop a rigorous test, we need to establish a distributional result for $\max_{k\le s_n} |\hat{\gamma}_k - \gamma_k|$, where s_n is the largest lag that can grow to infinity. It turns out that, with the physical dependence measure, we can formulate an asymptotic result for the maximum deviation $\max_{k\le s_n} |\hat{\gamma}_k - E\hat{\gamma}_k|$. Such a result can be used to construct simultaneous confidence intervals for γ_k with multiple lags. Let

$$\Delta_p(m) = \sum_{i=m}^{\infty} \delta_p(i), \ \Psi_p(m) = \left(\sum_{i=m}^{\infty} \delta_p^2(i)\right)^{1/2} \tag{15}$$

and

$$\Phi_p(m) = \sum_{i=0}^{\infty} \min\{\delta_p(i), \Psi_p(m)\}. \tag{16}$$

THEOREM 2. *(Xiao and Wu, 2011a) Assume that* $EX_i = 0$, $X_i \in \mathcal{L}^p$, $p > 4$, $\Delta_p(m) = O(m^{-\alpha})$, *and* $\Phi_p(m) = O(m^{-\alpha'})$, *where* $\alpha, \alpha' > 0$.

(i) If $\alpha > 1/2$ *or* $\alpha' p > 2$, *then for* $c_p = 6(p+4)e^{p/4}\Delta_4\|X_i\|_4$, *we have*

$$\lim_{n\to\infty} P\left(\max_{1\le k<n} |\hat{\gamma}_k - E\hat{\gamma}_k| \le c_p\sqrt{\frac{\log n}{n}}\right) = 1. \tag{17}$$

(ii) If $s_n \to \infty$ *satisfies* $s_n = O(n^\eta)$ *with* $0 < \eta < \min(1, \alpha p/2)$ *and* $\eta\min(2 - 4/p - 2\alpha, 1 - 2\alpha') < 1 - 4/p$, *then we have the Gumbel convergence: for all* $x \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left(\max_{1\le k\le s_n} \sqrt{n}|\hat{\gamma}_k - E\hat{\gamma}_k| \le \sigma_0^{1/2}(a_{2s_n}x + b_{2s_n})\right) = \exp(-\exp(-x)), \tag{18}$$

where $a_n = (2\log n)^{-1/2}$ *and* $b_n = a_n(4\log n - \log\log n - \log 4\pi)/2$.

3. Low-dimensional covariance matrix estimation

The problem of low-dimensional covariance matrix estimation often arises when one wants to estimate unknown parameters that are associated with a time series. Let θ_0 be an unknown parameter associated with the process (Y_i). Given observations $Y_1, \ldots, Y_n$, we estimate θ_0 by $\hat{\theta}_n = \hat{\theta}_n(Y_1, \ldots, Y_n)$. For example, if (Y_i) is a d-dimensional process with a common unknown mean vector $\mu_0 = EY_i$, then we can estimate it by the sample mean vector

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{19}$$

Under appropriate conditions on the process (Y_i), we expect that the central limit theorem for $\hat{\theta}_n$ holds:

$$\Sigma_n^{-1/2}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, \mathrm{Id}_d), \tag{20}$$

where Id_d is the d-dimensional identity matrix. With (20), one can construct confidence regions for θ_0. In particular, let $\hat{\Sigma}_n$ be an estimate of Σ_n. Then the $(1-\alpha)$th, $0 < \alpha < 1$, confidence ellipse for θ_0 is

$$\left\{\nu \in \mathbb{R}^d : (\hat{\theta}_n - \nu)^\top \hat{\Sigma}_n^{-1}(\hat{\theta}_n - \nu) = |\hat{\Sigma}_n^{-1/2}(\hat{\theta}_n - \nu)|^2 \le \chi^2_{d,1-\alpha}\right\}, \tag{21}$$

where $\chi^2_{d,1-\alpha}$ is the $(1-\alpha)$th quantile of a χ^2 distribution with degree of freedom d. The key question in the above construction now becomes the estimation of Σ_n. The latter question is closely related to the long-run variance estimation problem.

In the derivation of the central limit theorem (20), one typically needs to establish an asymptotic expansion of the type

$$\hat{\theta}_n - \theta_0 = \sum_{i=1}^{n} X_i + R_n, \tag{22}$$

where R_n is negligible in the sense that $\Sigma_n^{-1/2} R_n = o_{\mathbb{P}}(1)$ and (X_i) is a random process associated with (Y_i) satisfying the central limit theorem

$$\Sigma_n^{-1/2} \sum_{i=1}^{n} X_i \Rightarrow N(0, \mathrm{Id}_d).$$

Sometimes the expansion (22) is called the Bahadur (1966) representation. For i.i.d. random variables $Y_1, \ldots, Y_n$, Bahadur obtained an asymptotic linearizing approximation for its αth $(0 < \alpha < 1)$ sample quantile. Such an approximation greatly facilitates an asymptotic study. Note that the sample quantile depends on Y_i in a complicated nonlinear manner. The asymptotic expansion (22) can be obtained from the maximum likelihood, quasi maximum likelihood, or general method of moments estimation procedures. The random variables X_i in (22) are called scores or estimating functions. As another example, assume that (Y_i) is a stationary Markov process with transition density $p_{\theta_0}(Y_i|Y_{i-1})$, where θ_0 is an unknown parameter. Then given the observations $Y_0, \ldots, Y_n$, the conditional maximum likelihood estimate $\hat{\theta}_n$ maximizes

$$\ell_n(\theta) = \sum_{i=1}^{n} \log p_\theta(Y_i|Y_{i-1}). \tag{23}$$

As is common in the likelihood estimation theory, let $\dot{\ell}_n(\theta) = \partial \ell_n(\theta)/\partial\theta$ and $\ddot{\ell}_n(\theta) = \partial^2 \ell_n(\theta)/\partial\theta\partial\theta^\top$ be a $d \times d$ matrix. By the ergodic theorem, $\ddot{\ell}_n(\theta_0)/n \to E\ddot{\ell}_1(\theta_0)$ almost surely. Since $\dot{\ell}_n(\hat{\theta}_n) = 0$, under suitable conditions on the process (Y_i), we can perform the Taylor expansion $\dot{\ell}_n(\hat{\theta}_n) \approx \dot{\ell}_n(\theta_0) + \ddot{\ell}_n(\theta_0)(\hat{\theta}_n - \theta_0)$. Hence the representation (22) holds with

$$X_i = n^{-1}(E\ddot{\ell}_1(\theta_0))^{-1} \frac{\partial}{\partial\theta} \log p_\theta(Y_i|Y_{i-1})|_{\theta=\theta_0}. \tag{24}$$

A general theory for establishing (22) is presented in the study by Amemiya (1985) and Heyde (1997) and various special cases are considered in the study by Hall and Heyde (1980), Hall and Yao (2003), Wu (2007), He and Shao (1996), Klimko and Nelson (1978), Tong (1990), and among others.

For the sample mean estimate (19), it is also of form (22) by writing $\hat{\mu}_n - \mu_0 = n^{-1} \sum_{i=1}^{n} (Y_i - \mu_0)$ and $X_i = (Y_i - \mu_0)/n$. Therefore, to estimate the covariance matrix of an estimated parameter, in view of (22), we typically need to estimate the covariance matrix Σ_n of the sum $S_n = \sum_{i=1}^{n} X_i$. Clearly,

$$\Sigma_n = \sum_{1 \le i,j \le n} \mathrm{cov}(X_i, X_j), \tag{25}$$

where $\operatorname{cov}(X_i, X_j) = E(X_i X_j^\top) - E(X_i)E(X_j^\top)$. Sections 3.1, 3.2, and 3.3 concern convergence rates of estimates of Σ_n based on observations $(X_i)_{i=1}^n$, which can be independent, uncorrelated, nonstationary, and weakly dependent. In the estimation of the covariance matrix for $S_n = \sum_{i=1}^n X_i$ for $\hat{\theta}_n$ based on the representation (22), the estimating functions X_i may depend on the unknown parameter θ_0; hence, $X_i = X_i(\theta_0)$ may not be observed. For example, for the sample mean estimate (19), one has $X_i = (Y_i - \mu_0)/n$, while for the conditional MLE, X_i in (24) also depends on the unknown parameter θ_0. Heagerty and Lumley (2000) considered estimation of covariance matrices for estimated parameters for strong mixing processes; see also the study by Newey and West (1987) and Andrews (1991). In Corollary 1 of Section 3.2 and Section 3.4, we shall present asymptotic results for covariance matrix estimates with estimated parameters.

3.1. HC covariance matrix estimators

For independent but not necessarily identically distributed random vectors X_i, $1 \le i \le n$, White (1980) proposed a heteroscedasticity-consistent (HC) covariance matrix estimator for $\Sigma_n = \operatorname{var}(S_n)$, $S_n = \sum_{i=1}^n X_i$. Other contribution can be found in Eicker (1963) and MacKinnon and White (1985). If $\mu_0 = EX_i$ is known, we can estimate Σ_n by

$$\hat{\Sigma}_n^\circ = \sum_{i=1}^n (X_i - \mu_0)(X_i - \mu_0)^\top. \tag{26}$$

If μ_0 is unknown, we shall replace it by $\hat{\mu}_n = \sum_{i=1}^n X_i/n$ and form the estimate

$$\begin{aligned} \hat{\Sigma}_n &= \frac{n}{n-1} \sum_{i=1}^n (X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^\top \\ &= \frac{n}{n-1} \sum_{i=1}^n (X_i X_i^\top - \hat{\mu}_n \hat{\mu}_n^\top). \end{aligned} \tag{27}$$

Both $\hat{\Sigma}_n^\circ$ and $\hat{\Sigma}_n$ are unbiased for Σ_n. To this end, assume without loss of generality $\mu = 0$, then by independence, $n^2 E(\hat{\mu}_n \hat{\mu}_n^\top) = \sum_{i=1}^n E(X_i X_i^\top)$, hence

$$\begin{aligned} E\hat{\Sigma}_n &= \frac{n}{n-1} \left[\sum_{i=1}^n E(X_i X_i^\top) - E(n\hat{\mu}_n \hat{\mu}_n^\top) \right] \\ &= \sum_{i=1}^n E(X_i X_i^\top) = \Sigma_n. \end{aligned} \tag{28}$$

Theorem 3 below provides a convergence rate of $\hat{\Sigma}_n^\circ$. We omit its proof since it is an easy consequence of the Rothenthal inequality.

THEOREM 3. *Assume that X_i are independent $\mathbb{R}^d$ random vectors with $EX_i = 0, X_i \in \mathcal{L}^p$, $2 < p \le 4$. Then there exists a constant C, only depending on p and d, such that*

$$\|\hat{\Sigma}_n^\circ - \Sigma_n\|_{p/2}^{p/2} \le C \sum_{i=1}^n \|X_i\|_p^p. \tag{29}$$

As an immediate consequence of Theorem 3, if $\Sigma := \operatorname{cov}(X_i)$ does not depend on i and Σ is positive definite (namely $\Sigma > 0$) and $\sup_i \|X_i\|_p < \infty$, then

$$\|\hat{\Sigma}_n^\circ \Sigma_n^{-1} - \mathrm{Id}_d\|_{p/2} = O(n^{2/p-1})$$

and the confidence ellipse in (21) has an asymptotically correct coverage probability. Simple calculation shows that the above relation also holds if $\hat{\Sigma}_n^\circ$ is replaced by $\hat{\Sigma}_n$.

If X_i are uncorrelated, using the computation in (28), it is easily seen that the estimates $\hat{\Sigma}_n^\circ$ in (26) and $\hat{\Sigma}_n$ in (27) are still unbiased. However, one no longer has (29) if X_i are only uncorrelated instead of being independent. To establish an upper bound, as in Wu (2011), we assume that (X_i) has the form

$$X_i = H_i(\varepsilon_i, \varepsilon_{i-1}, \ldots), \tag{30}$$

where ε_i are i.i.d. random variables and H_i is a measurable function such that X_i is a proper random variable. If the function H_i does not depend on i, then (30) reduces to (5). In general, (30) defines a nonstationary process. According to the stochastic representation theory, any finite-dimensional random vector can be expressed in distribution as functions of i.i.d. uniform random variables; see the study by Wu (2011) for a review. As in (8), define the physical dependence measure

$$\delta_p(k) = \sup_i \|X_i - X_{i,k}\|_p, \quad k \ge 0, \tag{31}$$

where $X_{i,k}$ is a couple process of X_i with ε_{i-k} in the latter being replaced by ε'_{i-k}. For stationary processes of form (5), (8) and (31) are identical.

THEOREM 4. *Assume that X_i are uncorrelated with form (30) and $EX_i = 0, X_i \in \mathcal{L}^p$, $2 < p \le 4$. Let $\kappa_p = \sup_i \|X_i\|_p$. Then there exists a constant $C = C_{p,d}$ such that*

$$\|\hat{\Sigma}_n^\circ - \Sigma_n\|_{p/2} \le C n^{2/p} \kappa_p \sum_{k=0}^\infty \delta_p(k). \tag{32}$$

PROOF. Let $\alpha = p/2$. Since $X_i X_i^\top - EX_i X_i^\top = \sum_{k=0}^\infty \mathcal{P}_{i-k}(X_i X_i^\top)$ and $\mathcal{P}_{i-k}(X_i X_i^\top)$, $i = 1, \ldots, n$ are martingale differences, by the Burkholder and Minkowski inequalities, we have

$$\begin{aligned}\|\hat{\Sigma}_n^\circ - \Sigma_n\|_\alpha &\le \sum_{k=0}^\infty \Big\| \sum_{i=1}^n \mathcal{P}_{i-k}(X_i X_i^\top) \Big\|_\alpha \\ &\le C \sum_{k=0}^\infty \left[\sum_{i=1}^n \|\mathcal{P}_{i-k}(X_i X_i^\top)\|_\alpha^\alpha \right]^{1/\alpha}.\end{aligned}$$

Observe that $E[(X_{k,0}X_{k,0}^\top)|\mathcal{F}_0] = E[(X_kX_k^\top)|\mathcal{F}_{-1}]$. By Scharwz inequality, $\|\mathcal{P}_0(X_kX_k^\top)\|_\alpha \le \|X_{k,0}X_{k,0}^\top - X_kX_k^\top\|_\alpha \le 2\kappa_p\delta_p(k)$. Hence we have (32). □

3.2. *Long-run covariance matrix estimation for stationary vectors*

If X_i are correlated, then the estimate (27) is no longer consistent for Σ_n and autocovariances need to be taken into consideration. Recall $S_n = \sum_{i=1}^n X_i$. Assume $EX_i = 0$. Using the idea of lag window spectral density estimate, we estimate the covariance matrix $\Sigma_n = \mathrm{var}(S_n)$ by

$$\tilde{\Sigma}_n = \sum_{1\le i,j\le n} K\left(\frac{i-j}{B_n}\right) X_iX_j^\top, \tag{33}$$

where K is a window function satisfying $K(0) = 1$, $K(u) = 0$ if $|u| > 1$, K is even and differentiable on the interval $[-1, 1]$, and B_n is the lag sequence satisfying $B_n \to \infty$ and $B_n/n \to 0$. The former condition is for including unknown order of dependence, whereas the latter is for the purpose of consistency.

If (X_i) is a scalar process, then (33) is the lag-window estimate for the long-run variance $\sigma_\infty^2 = \sum_{k\in\mathbb{Z}} \gamma_i$, where $\gamma_i = \mathrm{cov}(X_0, X_i)$. Note that $\sigma_\infty^2/(2\pi)$ is the value of the spectral density of (X_i) at zero frequency. There is a huge literature on spectral density estimation; see the classical textbooks of Anderson (1971), Brillinger (1975), Brockwell and Davis (1991), Grenander and Rosenblatt (1957), Priestley (1981), and Rosenblatt (1985) and the third volume Handbook of Statistics "*Time Series in the Frequency Domain*" edited by Brillinger and Krishnaiah (1983). Rosenblatt (1985) showed the asymptotic normality for lag-window spectral density estimates for strong mixing processes under a summability condition of eighth-order joint cumulants.

Liu and Wu (2010) present an asymptotic theory for lag-window spectral density estimates under minimal moment and natural dependence conditions. Their results can be easily extended to the vector-valued processes. Assume $EX_i = 0$, then $\Sigma_n = \mathrm{var}(S_n)$ satisfies

$$\frac{1}{n}\Sigma_n = \sum_{k=1-n}^{n-1} (1 - |k|/n)E(X_0X_k^\top) \to \sum_{k=-\infty}^{\infty} E(X_0X_k^\top) =: \Sigma^\dagger. \tag{34}$$

Let vec be the vector operator. We have the following consistency and central limit theorem for $\mathrm{vec}(\tilde{\Sigma}_n)$. Its proof can be similarly carried out by using the argument in Liu and Wu (2010). Details are omitted.

THEOREM 5. *Assume that the d-dimensional stationary process (X_i) is of form (5), and $B_n \to \infty$ and $B_n = o(n)$. (i) If the short-range dependence condition (9) holds with $p \ge 2$, then $\|\tilde{\Sigma}_n/n - \Sigma_n/n\|_{p/2} = o(1)$ and, by (34), $\|\tilde{\Sigma}_n/n - \Sigma^\dagger\|_{p/2} = o(1)$. (ii) If (9) holds with $p = 4$, then there exists a matrix Γ with $\rho(\Gamma) < \infty$ such that*

$$(nB_n)^{-1/2}[\mathrm{vec}(\tilde{\Sigma}_n) - E\,\mathrm{vec}(\tilde{\Sigma}_n)] \Rightarrow N(0, \Gamma), \tag{35}$$

and the bias

$$n^{-1}\left\|E\operatorname{vec}(\tilde{\Sigma}_n)-\Sigma_n\right\| \le \sum_{k=-B_n}^{B_n}|1-K(k/B_n)|\,\gamma_2(k)+2\sum_{k=B_n+1}^{n}\gamma_2(k), \tag{36}$$

where $\gamma_2(k)=\left\|E(X_0X_{i+k}^{\top})\right\| \le \sum_{i=0}^{\infty}\delta_2(i)\delta_2(i+k)$.

An interesting feature of Theorem 5(i) is that, under the minimal moment condition $X_i \in \mathcal{L}^2$ and the very mild weak dependence condition $\Delta_2 < \infty$, the estimate $\tilde{\Sigma}_n/n$ is consistent for Σ_n/n. This property substantially extends the range of applicability of lag-window covariance matrix estimates. For consistency, the study by Andrews (1991) requires a finite fourth moment and a fourth-order joint cumulant summability condition, while for computing the asymptotic mean square error, it needs a finite eighth moment and an eighth-order cumulant summability condition. For nonlinear processes, it might be difficult to verify those cumulant summability conditions. Our framework of physical dependence measure seems quite convenient and useful for long-run covariance matrix estimation, and it is no longer needed to work with joint cumulants.

In many situations, X_i depends on unknown parameters and thus is not directly observable. For example, X_i in (24) depends on the unknown parameter θ_n. Then it is natural to modify the $\tilde{\Sigma}_n$ in (33) by the following estimate

$$\tilde{\Sigma}_n(\hat{\theta}_n)=\sum_{1\le i,j\le n}K\left(\frac{i-j}{B_n}\right)X_i(\hat{\theta}_n)X_j(\hat{\theta}_n)^{\top}, \tag{37}$$

where $\hat{\theta}_n$ is an estimate of θ_0, so that $X_i(\hat{\theta}_n)$ are estimates of $X_i(\theta_0)=X_i$. Note that $\tilde{\Sigma}_n(\theta_0)=\tilde{\Sigma}_n$. As in the study by Newey and West (1987) and Andrews (1991), appropriate continuity conditions on the random function $X_i(\cdot)$ can imply the consistency of the estimate $\tilde{\Sigma}_n(\hat{\theta}_n)$. The following Corollary 1 is a straightforward consequence of Theorem 5.

COROLLARY 1. *Assume that* $\hat{\theta}_n$ *is a* $\sqrt{n}$*-consistent estimate of* θ_0*, namely* $\sqrt{n}(\hat{\theta}_n-\theta_0)=O_{\mathbb{P}}(1)$*. Further assume that there exists a constant* $\delta_0>0$ *such that the local maximal function* $X_i^*=\sup\{|\partial X_i(\theta)/\partial\theta| : |\theta-\theta_0|\le\delta_0\}\in\mathcal{L}^2$*. Assume* $B_n\to\infty$*,* $B_n=o(\sqrt{n})$ *and (9) holds with* $p=2$*. Then* $\tilde{\Sigma}_n(\hat{\theta}_n)/n-\Sigma_n/n\to 0$ *in probability.*

3.3. HAC covariance matrix estimators

Recall (31) for the definition of physical dependence measures for nonstationary processes. If (X_i) is nonstationary and correlated, then under a similar short-range dependence condition as (9), we can also obtain a convergence rate for the estimate $\tilde{\Sigma}_n$ defined in (33). Results of similar type were given in the study by Newey and West (1987) and Andrews (1991). Andrews and Monahan (1992) improved the estimate by using a prewhitening procedure.

THEOREM 6. *Let $B_n \to \infty$ and $B_n/n \to 0$. Assume that the nonstationary process (X_i) is of form (30), $X_i \in \mathcal{L}^p$, $p > 2$, and the short-range dependence condition (9). (i) If $2 < p < 4$, then $\|\tilde{\Sigma}_n/n - \Sigma_n/n\|_{p/2} = o(1)$. (ii) If $p \geq 4$, then there exists a constant C depending on p and d only such that*

$$\|\tilde{\Sigma}_n - E\tilde{\Sigma}_n\|_{p/2} \leq C\Delta_p^2 B_n, \tag{38}$$

and the bias

$$n^{-1}\left\|E\tilde{\Sigma}_n - \Sigma_n\right\| \leq \sum_{k=-B_n}^{B_n} |1 - K(k/B_n)|\,\gamma_2(k) + 2\sum_{k=B_n+1}^{n} \gamma_2(k), \tag{39}$$

where $\gamma_2(k) = \sup_i \left\|E(X_i X_{i+k}^\top)\right\| \leq \sum_{i=0}^{\infty} \delta_2(i)\delta_2(i+k)$.

REMARK 1. We emphasize that in this section, since the dimension of the covariance matrix Σ is fixed, all matrix norms are essentially equivalent and the relations (29), (32), (36), and (38) also hold if we use other types of matrix norms such as the Frobenius norm and the maximum entry norm. This feature is no longer present for high-dimensional matrix estimation where the dimension can be unbounded; see Section 4. □

As Theorem 5, the proof of Theorem 6 can be similarly carried out by using the argument in the study by Liu and Wu (2010). A crucial step in applying Theorem 6 is how to choose the smoothing parameter. See the study by Zeileis (2004) for an excellent account for the latter problem.

3.4. Covariance matrix estimation for linear models

Consider the linear model

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + e_i, \quad 1 \leq i \leq n, \tag{40}$$

where $\boldsymbol{\beta}$ is an $s \times 1$ unknown regression coefficient vector, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{is})'$ are $s \times 1$ known (nonstochastic) design vectors. Let $\hat{\boldsymbol{\beta}}$ be the least square estimate of $\boldsymbol{\beta}$. Here we consider the estimation of $\operatorname{cov}(\hat{\boldsymbol{\beta}})$ under the assumption that (e_i) is a nonstationary process of form (30). As a special case, if there is only one covariate and $x_i = 1$ for each $1 \leq i \leq n$, then $\hat{\boldsymbol{\beta}} = S_n/n$ with $S_n = \sum_{i=1}^n e_i$, so the estimation of the covariance matrix of S_n in Section 3.3 is a special case here. Assume that for large n, $T_n := \mathbf{X}_n^\top \mathbf{X}_n$ is positive definite. It is more convenient to consider the rescaled model

$$y_i = \boldsymbol{z}_i^\top \boldsymbol{\theta} + e_i \quad \text{with} \quad \boldsymbol{z}_i = \boldsymbol{z}_{i,n} = T_n^{-1/2}\boldsymbol{x}_i \text{ and } \boldsymbol{\theta} = \boldsymbol{\theta}_n = T_n^{1/2}\boldsymbol{\beta}, \tag{41}$$

under which the least square estimate $\hat{\boldsymbol{\theta}} = \sum_{i=1}^n \boldsymbol{z}_i e_i$. If e_i were known, we can estimate $\Sigma_n := \operatorname{cov}(\hat{\boldsymbol{\theta}})$ by

$$V_n = \sum_{1 \leq i,j \leq n} K\left(\frac{i-j}{B_n}\right) \mathbf{z}_i e_i \mathbf{z}_j^\top e_j, \tag{42}$$

which is in the similar fashion as (33). Since e_i are unknown, we should replace e_i in V_n by the estimated residuals $\hat{e}_i$ and employ the following estimate

$$\hat{V}_n = \sum_{1\le i,j\le n} z_i \hat{e}_i z_j^\top \hat{e}_j c_{ij}, \tag{43}$$

where $c_{ij} = K((i-j)/B_n)$. We have the following convergence rate of the estimate $\hat{V}_n$, which can be derived using similar arguments as those in the study by Liu and Wu (2010).

THEOREM 7. *Assume that the nonstationary process (e_i) is of form (30), $e_i \in \mathcal{L}^p$ with $p \ge 4$, and $\Delta_p < \infty$. Let $c_k := K(k/B_n)$. Then there exists a constant C depending only on p and s such that*

$$\left\| \hat{V}_n - EV_n \right\|_{p/2} \le C\Delta_p^2 \left(\sum_{1\le i,j\le n} c_{i-j}^2 |z_i|^2 |z_j|^2 \right)^{1/2}, \tag{44}$$

and the bias

$$\|EV_n - \Sigma_n\| \le s \sum_{k=1-n}^{n-1} |1 - c_k| \gamma_2(k), \tag{45}$$

where $\gamma_2(k) = \sup_{i\in\mathbb{Z}} |E(e_i e_{i+k})| \le \sum_{i=0}^{\infty} \delta_2(i)\delta_2(i+|k|)$.

In Example 1 of Section 4.4, we shall obtain a best linear unbiased estimate for β by estimating the high-dimensional covariance matrix of $(e_1, \ldots, e_n)$. It illustrates the different natures of two types of covariance matrix estimation.

4. High-dimensional covariance matrix estimation

In this section, we shall consider estimation of high-dimensional covariance matrices in time series in which the dimensions can grow to infinity. This setting is quite different from the one in (25), where the dimension is fixed and does not grow. During the last decade, the problem of high-dimensional covariance matrix estimation has attracted considerable attention. See the work done by Pourahmadi (2011) for an excellent review. The problem is quite challenging since, for estimating Σ_p given in (1), one has to estimate $p(p+1)/2$ unknown parameters. Additionally, those parameters must follow the highly nontrivial positive-definiteness constraint. In the multivariate setting in which one has multiple i.i.d. p-variate random variables, the problem has been extensively studied; see the wok done by Meinshausen and Bühlman (2006), Yuan and Lin (2007), Rothman et al. (2009), Bickel and Levina (2008a,b), Cai et al. (2010), Lam and Fan (2009), Ledoit and Wolf (2004) and among others. As commented in the study by Bickel and Gel (2011), the same problem in longitudinal and time series setting has been much less investigated. In comparison with the matrix estimation problem in

the context of multivariate statistics, there are several distinguished features when we consider time series:

(i) The order information among observations is highly relevant;
(ii) Variables that are far apart are weakly dependent;
(iii) The number of replicates is very small, and in many cases, there is only one realization available.

In multivariate statistics, in many cases, one can permute the variables without sacrificing the interpretability, and the permutation-invariance property of a covariance matrix estimate can be quite appealing. For covariance matrix estimation in time series, however, the permutation-invariance property is not a must.

Section 4.1 reviews the Cholesky decomposition of covariance matrices. As argued in the study by Pourahmadi (1999), the Cholesky decomposition-based covariance matrix estimate is inherently positive definite and its entries have a nice interpretation of being autoregressive coefficients. In Section 4.2, we briefly review parametric covariance matrix estimation where the target covariance matrix is of certain parametric forms. Thus, it suffices to estimate the governing parameters. Sections 4.3 and 4.4 concern the nonparametric covariance matrix estimation problem for two different settings: in the first setting, we assume that there are multiple i.i.d. realizations of the underlying process, whereas in the second one only one realization is available. For the latter, we assume that the underlying process is stationary.

4.1. *Cholesky decomposition*

Assume that $X_1, \ldots, X_p$ is a mean zero Gaussian process with covariance matrix Σ_p given in (1). As in the study by Pourahmadi (1999), we perform successive autoregression of X_t on its predecessors $X_1, \ldots, X_{t-1}$ in the following manner:

$$X_t = \sum_{j=1}^{t-1} \phi_{tj} X_j + \eta_t =: \hat{X}_t + \eta_t, \ t = 1, \ldots, p, \tag{46}$$

where ϕ_{tj} are the autoregressive coefficients such that $\hat{X}_t$ is the projection of X_t onto the linear space spanned by $X_1, \ldots, X_{t-1}$. Then $\eta_1 \equiv X_1$ and $\eta_t = X_t - \hat{X}_t$, $t = 2, \ldots, n$, are independent. Let $\sigma_t^2 = \mathrm{var}(\eta_t)$ be the innovation variance, $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ and

$$L = \begin{pmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{31} & -\phi_{32} & 1 & & \\ \cdots & \cdots & \cdots & \cdots & \\ -\phi_{11} & -\phi_{p2} & \cdots & -\phi_{p,p-1} & 1 \end{pmatrix} \tag{47}$$

be a lower triangle matrix. Then Σ_p has the representation

$$L \Sigma_p L^\top = D^2, \tag{48}$$

which implies the useful fact that the inverse, or the precision matrix,

$$\Sigma_p^{-1} = LD^{-2}L^\top. \tag{49}$$

An important feature of the representation (48) is that the coefficients in L are unconstrained, and if an estimate of Σ_p is computed based on estimated L and D, then it is guaranteed to be non-negative definite. The Cholesky method is particularly suited for covariance and precision matrix estimation in time series, and the entries in L can be interpreted as autoregressive coefficients.

Another popular method is the eigen decomposition $\Sigma_p = Q\Lambda Q^\top$, where Q is an orthonormal matrix, namely $QQ^\top = \mathrm{Id}_p$ and Λ is a diagonal matrix that consists of eigenvalues of Σ_p. The eigen decomposition is related to the principal component analysis. It is generally not easy to work with the orthonormality constraint. See the work done by Pourahmadi (2011) for more discussion.

4.2. *Parametric covariance matrix estimation*

In the parametric covariance matrix estimation problem, one assumes that Σ_n has a known form $\Sigma_n(\theta)$ indexed by a finite-dimensional parameter. To estimate Σ_n, it would then suffice if we can find a good estimate of θ. Anderson (1970) assumed that Σ_n is a linear combination of some known matrices. Burg et al. (1982) applied the maximum likelihood estimation method; see also the study by Quang (1984), Dembo (1986), Fuhrmann and Miller (1988), Jansson and Ottersten (2000), and Dietrich (2008). Chiu et al. (1996) used a log-linear covariance matrix parametrization.

Based on the Cholesky decomposition (48), Pourahmadi (1999) considered parametric modelling for the autoregressive coefficients ϕ_{ij} and the innovation variance σ_i^2, thus substantially reducing the number of parameters. See also the study by Pan and MacKenzie (2003) and Zimmerman and Núñez-Antón (2010).

4.3. *Covariance matrix estimation with multiple i.i.d. realizations*

Assume that $(X_{l,1}, X_{l,2}, \ldots, X_{l,p})$, $l = 1, \ldots, m$, are i.i.d. random vectors identically distributed as $(X_1, \ldots, X_p)$. If the means $\mu_j = EX_{l,j}$, $j = 1, \ldots, p$, are known, then the covariance $\gamma_{i,j} = \mathrm{cov}(X_{l,i}, X_{l,j})$, $1 \le i, j \le p$, can be estimated by

$$\hat{\gamma}_{i,j} = \frac{1}{m}\sum_{l=1}^{m}(X_{l,i} - \mu_i)(X_{l,j} - \mu_j), \tag{50}$$

and the sample covariance matrix estimate is

$$\hat{\Sigma}_p = (\hat{\gamma}_{i,j})_{1\le i,j\le p}. \tag{51}$$

If μ_j is unknown, one can naturally estimate it by the sample mean $\bar{\mu}_j = m^{-1}\sum_{l=1}^{m} X_{l,j}$ and $\hat{\gamma}_{i,j}$ and $\hat{\Sigma}_p$ in (50) and (51) can then be modified correspondingly.

According to the modern random matrix theory, under the assumption that all entries $X_{l,i}, 1 \le l \le m, 1 \le i \le p$, are independent, $\hat{\Sigma}_p$ is a bad estimate of Σ_p in the sense

that it is inconsistent in operator norm. Such inconsistency results for sample covariance matrices in multivariate analysis have been discussed in the study by Stein (1975), Bai and Silverstein (2010), El Karoui (2007), Paul (2007), Johnstone (2001), Geman (1980), Wachter (1978), Anderson et al. (2010), and among others. Note that if $m < p$, $\hat{\Sigma}_p$ is a singular matrix. It is known that, under appropriate moment conditions of $X_{l,i}$, if $p/m \to c$, then the empirical distribution of eigenvalues of $\hat{\Sigma}_p$ follows the Marcenko–Pastur law that has the support $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$ and a point mass at zero if $c > 1$; and the largest eigenvalue, after proper normalization, follows the Tracy–Widom law. All those results suggest the inconsistency of sample covariance matrices.

For an improved and consistent estimation, various regularization methods have been proposed. Assuming that the correlations are weak if the lag $i - j$ is large, Bickel and Levina (2008a) proposed the banded covariance matrix estimate

$$\hat{\Sigma}_{p,B} = (\hat{\gamma}_{i,j} \mathbf{1}_{|i-j| \le B})_{1 \le i,j \le p}, \tag{52}$$

where $B = B_p$ is the band parameter, and more generally, the tapered estimate

$$\hat{\Sigma}_{p,B} = (\hat{\gamma}_{i,j} K(|i-j|/B))_{1 \le i,j \le p}, \tag{53}$$

where K is a symmetric window function with support on $[-1, 1]$, $K(0) = 1$, and K is continuous on $(-1, 1)$. Here we assume that $B_p \to \infty$ and $B_p/p \to 0$. The former condition ensures that $\hat{\Sigma}_{p,B}$ can include dependencies at unknown orders, whereas the latter aims to circumvent the weak signal-to-noise ratio issue that $\hat{\gamma}_{i,j}$ is a bad estimate of $\gamma_{i,j}$ if $|i-j|$ is big. In particular, Bickel and Levina (2008a) considered the class

$$\mathcal{U}(\epsilon_0, \alpha, C) = \left\{ \Sigma : \max_j \sum_{i:\,|i-j|>k} |\gamma_{i,j}| \le C k^{-\alpha},\ \rho(\Sigma) \le \epsilon_0^{-1}, \rho(\Sigma^{-1}) \le \epsilon_0 \right\}. \tag{54}$$

This condition quantifies issue (ii) mentioned in the beginning of this section. They proved that (i) if $\max_j E \exp(u X_{l,i}^2) < \infty$ for some $u > 0$ and $k_n \asymp (m^{-1} \log p)^{-1/(2\alpha+2)}$, then

$$\rho(\hat{\Sigma}_{p,k_p} - \Sigma_p) = \mathcal{O}_P[(m^{-1} \log p)^{\alpha/(2\alpha+2)}]; \tag{55}$$

(ii) if $\max_j E|X_{l,i}|^\beta < \infty$ and $k_n \asymp (m^{-1/2} p^{2/\beta})^{c(\alpha)}$, where $c(\alpha) = (1 + \alpha + 2/\beta)^{-1}$, then

$$\rho(\hat{\Sigma}_{p,k_p} - \Sigma_p) = \mathcal{O}_P[(m^{-1/2} p^{2/\beta})^{\alpha c(\alpha)}]. \tag{56}$$

In the tapered estimate (53), if we choose K such that the matrix $W_p = (K(|i-j|/l))_{1 \le i,j \le p}$ is positive definite, then $\tilde{\Sigma}_{p,l}$ is the Hadamard (or Schur) product of $\hat{\Sigma}_n$ and W_p, and by the Schur Product Theorem in matrix theory (Horn and Johnson, 1990), it is also non-negative definite since $\hat{\Sigma}_n$ is non-negative definite. For example, W_n is positive definite for the triangular window $K(u) = \max(0, 1 - |u|)$ or the

Parzen window $K(u) = 1 - 6u^2 + 6|u|^3$ if $|u| < 1/2$ and $K(u) = \max[0, 2(1 - |u|)^3]$ if $|u| \geq 1/2$.

Based on the Cholesky decomposition (48), Wu and Pourahmadi (2003) proposed a nonparametric estimator for the precision matrix Σ_p^{-1} for locally stationary processes (Dahlhaus, 1997), which are time-varying AR processes

$$X_t = \sum_{j=1}^{k} f_j(t/p) X_{t-j} + \sigma(t/p)\eta_t^0. \tag{57}$$

Here η_t^0 are i.i.d. random variables with mean 0 and variance 1, and $f_j(\cdot)$ and $\sigma(\cdot)$ are continuous functions. Hence $\phi_{t,t-j} = f_j(t/p)$ if $1 \leq j \leq k$ and $\phi_{t,t-j} = 0$ if $j > k$. Wu and Pourahmadi (2003) applied a two-step method for estimating $f_j(\cdot)$ and $\sigma(\cdot)$: the first step is that, based on the data $(X_{l,1}, X_{l,2}, \ldots, X_{l,p})$, $l = 1, \ldots, m$, we perform a successive linear regression and obtain the least squares estimate $\hat{\phi}_{t,t-j}$ and the prediction variance $\hat{\sigma}^2(t/p)$; in the second step, we do a local linear regression on the raw estimates $\hat{\phi}_{t,t-j}$ and obtain smoothed estimates $\hat{f}_j(\cdot)$. Then we piece those estimates together and obtain an estimate for the precision matrix Σ_p^{-1} by (49). The lag k can be chosen by AIC, BIC, or other information criteria. Huang et al. (2006) applied a penalized likelihood estimator that is related to LASSO and ridge regression.

4.4. *Covariance matrix estimation with one realization*

If there is only one realization available, then it is necessary to impose appropriate structural assumptions on the underlying process and otherwise it would not be possible to estimate its covariance matrix. Here we shall assume that the process is stationary; hence, Σ_n is Toeplitz and $\gamma_{i,j} = \gamma_{i-j}$ can be estimated by the sample autocovariance (3) or (4), depending on whether the mean μ is known or not.

Covariance matrix estimation of stationary processes has been widely studied in the engineering literature. Lifanov and Likharev (1983) performed maximum likelihood estimation with applications in radio engineering. Christensen (2007) applied an EM-algorithm for estimating band-Toeplitz covariance matrices. Other contributions for estimating Toeplitz covariance matrices can be found in the study by Jansson and Ottersten (2000) and Burg et al. (1982). See also Chapter 3 in the excellent monograph of Dietrich (2008). However, in most of those papers, it is assumed that multiple i.i.d. realizations are available.

For a stationary process (X_i), Wu and Pourahmadi (2009) proved that the sample autocovariance matrix $\hat{\Sigma}_p$ is not a consistent estimate of Σ_p. A refined result is obtained in the study by Xiao and Wu (2011b) and they derived the exact order of $\rho(\hat{\Sigma}_p - \Sigma_p)$.

THEOREM 8. *(Xiao and Wu, 2011b). Assume that $X_i \in \mathcal{L}^\beta$, $\beta > 2$, $EX_i = 0$, $\Delta_\beta(m) = o(1/\log m)$, and $\min_\theta f(\theta) > 0$. Then*

$$\lim_{n\to\infty} P\left[\frac{\pi \min_\theta f^2(\theta)}{12\Delta_2^2} \log p \leq \rho(\hat{\Sigma}_p) \leq 10\Delta_2^2 \log p\right] = 1. \tag{58}$$

To obtain a consistent estimate of Σ_p, following the idea of lag-window spectral density estimation and tapering, we define the tapered covariance matrix estimate

$$\hat{\Sigma}_{p,B} = \left[K((i-j)/B)\hat{\gamma}_{i-j}\right]_{1\le i,j\le p} = \hat{\Sigma}_p \star W_p, \tag{59}$$

where $B = B_p$ is the bandwidth satisfying $B_p \to \infty$ and $B_p/p \to 0$, and $K(\cdot)$ is a symmetric kernel function with

$$K(0) = 1, \quad |K(x)| \le 1, \quad \text{and} \quad K(x) = 0 \text{ for } |x| > 1. \tag{60}$$

Estimate (59) has the same form as Bickel and Levina's (52) with the sample covariance matrix replaced by the sample autocovariance matrix. The form (59) is also considered in the study by McMurry and Politis (2010). Toeplitz (1911) studied the infinite-dimensional matrix $\Sigma_\infty = (a_{i-j})_{i,j\in\mathbb{Z}}$ and proved that its eigenvalues coincide with the image set $\{g(\theta) : \theta \in [0, 2\pi)\}$, where

$$g(\theta) = \sum_{j\in\mathbb{Z}} a_j e^{\sqrt{-1}j\theta}. \tag{61}$$

Note that $2\pi g(\theta)$ is the Fourier transform of (a_j). For a finite $p \times p$ matrix $\Sigma_p = (a_{i-j})_{1\le i,j\le p}$, its eigenvalues are approximately equally distributed as $\{g(\theta_j),\ j = 0, \ldots, p-1\}$, where $\theta_j = 2\pi j/p$ are the Fourier frequencies. See the excellent monograph by Grenander and Szegö (1958) for a detailed account. Hence the eigenvalues of the matrix estimate $\hat{\Sigma}_{p,B}$ in (59) are expected to be close to the image set of the lag-window estimate

$$\hat{f}_{p,B}(\theta) = \frac{1}{2\pi}\sum_{k=-B}^{B} K(k/B)\hat{\gamma}_k \cos(k\theta). \tag{62}$$

Using an asymptotic theory for lag-window spectral density estimates, Xiao and Wu (2011b) derived a convergence rate for $\rho(\hat{\Sigma}_{p,B} - \Sigma_p)$. Recall (15) and (16) for $\Delta_p(m)$ and $\Phi_p(m)$.

THEOREM 9. *(Xiao and Wu, 2011b) Assume $X_i \in \mathcal{L}^\beta$, $\beta > 4$, $EX_i = 0$, and $\Delta_p(m) = O(m^{-\alpha})$. Assume $B \to \infty$ and $B = O(p^\gamma)$, where $0 < \gamma < \min(1, \alpha\beta/2)$ and $(1 - 2\alpha)\gamma < 1 - 4/\beta$. Let $c_\beta = (\beta + 4)e^{\beta/4}$. Then*

$$\lim_{n\to\infty} P\left[\rho(\hat{\Sigma}_{p,B} - E\hat{\Sigma}_{p,B}) \le 12c_\beta\Delta_4^2\sqrt{\frac{B\log B}{p}}\right] = 1. \tag{63}$$

In particular, if $K(x) = \mathbf{1}_{\{|x|\le 1\}}$ is the rectangular kernel and $B \asymp (p/\log p)^{1/(2\alpha+1)}$, then

$$\rho(\hat{\Sigma}_{p,B} - \Sigma_p) = O_P\left[\left(\frac{\log p}{p}\right)^{\frac{\alpha}{2\alpha+1}}\right]. \tag{64}$$

The uniform convergence result in Theorem 2 motivates the following thresholded estimate:

$$\hat{\Sigma}_{p,T}^{\ddagger} = (\hat{\gamma}_{i-j}\mathbf{1}_{|\hat{\gamma}_{i-j}|\geq T})_{1\leq i,j\leq p}. \tag{65}$$

It is a shrinkage estimator. Note that $\hat{\Sigma}_{p,T}^{\ddagger}$ may not be positive. Bickel and Levina (2008b) considered the above estimate under the assumption that one has multiple i.i.d. realizations.

THEOREM 10. *(Xiao and Wu, 2011b) Assume* $X_i \in \mathcal{L}^\beta$, $\beta > 4$, $EX_i = 0$, $\Delta_p(m) = O(m^{-\alpha})$, *and* $\Phi_p(m) = O(m^{-\alpha'})$, $\alpha \geq \alpha' > 0$. *Let* $T = 6c_\beta \|X_0\|_4 \Delta_2 \sqrt{p^{-1}\log p}$. *If* $\alpha > 1/2$ *or* $\alpha'\beta > 2$, *then*

$$\rho\left(\hat{\Sigma}_{p,T}^{\ddagger} - \Sigma_p\right) = O_P\left[\left(\frac{\log p}{p}\right)^{\frac{\alpha}{2\alpha+2}}\right]. \tag{66}$$

Example 1. Here we shall show how to obtain a BLUE (best linear unbiased estimate) for linear models with dependent errors. Consider the linear regression model (40)

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + e_i, \quad 1 \leq i \leq p, \tag{67}$$

where now we assume that (e_i) is stationary. If the covariance matrix Σ_p of $(e_1, \ldots, e_p)$ is known, then the BLUE for $\boldsymbol{\beta}$ is of the form

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \Sigma_p^{-1}\mathbf{X})^{-1}\Sigma_p^{-1/2}\boldsymbol{y}, \tag{68}$$

where $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$ and $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)^\top$. If Σ_p is unknown, we estimate $\boldsymbol{\beta}$ by a two-step method. Using the ordinary least squares approach, we obtain a preliminary estimate $\bar{\boldsymbol{\beta}}$ and compute the estimated residuals $\hat{e}_i = y_i - \boldsymbol{x}_i^T\bar{\boldsymbol{\beta}}$. Based on the latter, using the tapered estimate $\tilde{\Sigma}_p$ of form (59) for Σ_p, a refined estimate of $\tilde{\boldsymbol{\beta}}$ can be obtained via (68) by using the weighted least squares with the weight matrix $\tilde{\Sigma}_p$. Due to the consistency of $\tilde{\Sigma}_p$, the resulting estimate for β is asymptotically BLUE.

Acknowledgments

This work was supported in part from DMS-0906073 and DMS-1106970. We thank reviewer for his/her comments that lead to an improved version.

References

Amemiya, T., 1985. Advanced Econometrics. Cambridge, Harvard University Press.

Anderson, G.W., Guionnet, A., Zeitouni, O., 2010. An introduction to random matrices, Cambridge Studies in Advanced Mathematics, 118, Cambridge University Press, Cambridge.

Anderson, T.W., 1970. Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. In: Essays in Probability and Statistics, pp. 1–24. The University of North Carolina Press, Chapel Hill.

Anderson, T.W., 1971. The Statistical Analysis of Time Series. Wiley, New York.

Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858.

Andrews, D.W.K., Monahan, J.C., 1992. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. Econometrica 60, 953–966.

Bai, Z., Silverstein, J.W., 2010. Spectral Analysis of Large Dimensional Random Matrices, second ed. Springer, New York.

Bahadur, R.R., 1966. A note on quantiles in large samples. Ann. Math. Stat. 37, 577–580.

Bartlett, M.S., 1946. On the theoretical specification and sampling properties of autocorrelated time-series. Suppl. J. Roy. Stat. Soc. 8, 27–41.

Bickel, P.J., Levina, E., 2008a. Regularized estimation of large covariance matrices. Ann. Stat. 36, 199–227.

Bickel, P.J., Levina, E., 2008b. Covariance regularization by thresholding. Ann. Stat. 36, 2577–2604.

Bickel P., Gel, Y., 2011. Banded regularization of covariance matrices in application to parameter estimation and forecasting of time series. J. Roy. Stat. Soc. B. 73, 711–728.

Borkar, V.S., 1993. White-noise representations in stochastic realization theory. SIAM J. Control Optim. 31, 1093–1102.

Bradley, R.C., 2007. Introduction to Strong Mixing Conditions. Kendrick Press, Utah.

Brillinger, D.R., 1975. Time series. Data analysis and theory. International Series in Decision Processes. Holt, Rinehart and Winston, Inc., New York-Montreal, London.

Brillinger, D.R., Krishnaiah, P.R. (Eds.), 1983. Handbook of Statistics 3: Time Series in the Frequency Domain, North-Holland Publishing Co., Amsterdam.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods, second ed. Springer, New York.

Burg, J.P., Luenberger, D.G., Wenger, D.L., 1982. Estimation of structured covariance matrices. Proc. IEEE 70, 963–974.

Cai, T., Zhang, C.H., Zhou, H., 2010. Optimal rates of convergence for covariance matrix estimation. Ann. Stat. 38, 2118–2144.

Chiu, T.Y.M., Leonard, T., Tsui, K.W., 1996. The matrix-logarithm covariance model. J. Amer. Stat. Assoc. 91, 198–210.

Christensen, L.P.B., 2007. An EM-algorithm for Band-Toeplitz Covariance Matrix Estimation. In: IEEE International Conference on Acoustics, Speech and Signal Processing III, Honolulu, pp. 1021–1024.

Dahlhaus, R., 1997. Fitting time series models to nonstationary processes. Ann. Stat. 36, 1–37.

de Jong, R.M., Davidson, J., 2000. Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. Econometrica 68, 407–423.

Dembo, A., 1986. The relation between maximum likelihood estimation of structured covariance matrices and periodograms. IEEE Trans. Acoust., Speech, Signal Processing 34(6), 1661–1662.

Dietrich, F.A., 2008. Robust Signal Processing for Wireless Communications. Springer, Berlin.

Doukhan, P., 1994. Mixing: Properties and Examples. Springer, New York.

Eberlein, E., Taqqu, M., (Ed.), 1986. Dependence in Probability and Statistics: A Survey of Recent Results. Birkhauser, Boston.

Eicker, F., 1963. Asymptotic normality and consistency of the least squares estimator for families of linear regressions. Ann. Math. Stat. 34, 447–456.

El Karoui, N., 2007. Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. Ann. Probab. 35, 663–714.

Fuhrmann, D.R., Miller, M.I., 1988. On the existence of positive-definite maximum-likelihood estimates of structured covariance matrices. IEEE Trans. Inform. Theor. 34(4), 722–729.

Geman, S., 1980. A limit theorem for the norm of random matrices. Ann. Probab. 8, 252–261.

Grenander, U., Rosenblatt, M., 1957. Statistical Analysis of Stationary Time Series. Wiley, New York.

Grenander, U., Szegö, G., 1958. Toeplitz Forms and Their Applications. Berkeley, CA, University of California Press.

Hall, P., Heyde, C.C., 1980. Martingale Limit Theorem and its Application. Academic Press, New York.

Hall, P., Yao, Q.W., 2003. Inference in ARCH and GARCH models with heavy-tailed errors. Econometrica 71, 285–317.

Hannan, E.J., 1970. Multiple Time Series. Wiley, New York.

Hannan, E.J., 1976. The asymptotic distribution of serial covariances. Ann. Stat. 4, 396–399.
Harris, D., McCabe, B., Leybourne, S., 2003. Some limit theory for autocovariances whose order depends on sample size. Economet. Theor. 19, 829–864.
He, X., Shao, Q.-M., 1996. A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. Ann. Stat. 24, 2608–2630.
Heagerty, P.J., Lumley, T., 2000. Window subsampling of estimating functions with application to regression models. J. Amer. Stat. Assoc. 95, 197–211.
Heyde, C.C., 1997. Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation, Springer, New York.
Horn, R.A., Johnson, C.R., 1990. Matrix Analysis. Corrected reprint of the 1985 original. Cambridge University Press, Cambridge, UK.
Hosking, J.R.M., 1996. Asymptotic distributions of the samplemean, autocovariances, and autocorrelations of long-memory timeseries. J. Econom. 73, 261–284.
Huang, J.Z., Liu, N., Pourahmadi, M., Liu, L., 2006. Covariance matrix selection and estimation via penalised normal likelihood. Biometrika 93, 85–98.
Ibragimov, I.A., Linnik, Y.V., 1971. Independent and Stationary Sequences of Random Variables. Groningen, Wolters-Noordhoff.
Jansson, M., Ottersten, B., 2000. Structured covariance matrix estimation: A parametric approach, 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing 5, 3172–3175.
Johnstone, I.M., 2001. On the distribution of the largest eigenvalue in principal components analysis. Ann. Stat. 29, 295–327.
Kalikow, S.A., 1982. T, T^{-1} transformation is not loosely Bernoulli. Ann. Math. 115, 393–409.
Kallianpur, G., 1981. Some ramifications of Wieners ideas on nonlinear prediction. In: Norbert Wiener, Collected Works with Commentaries. MIT Press, Mass., pp. 402–424.
Keenan, D.M., 1997. A central limit theorem for m(n) autocovariances. J. Time Ser. Anal. 18, 61–78.
Klimko, L.A., Nelson, P.I., 1978. On conditional least squares estimation for stochastic processes. Ann. Stat. 6, 629–642.
Lam, C., Fan, J., 2009. Sparsistency and rates of convergence in large covariance matrix estimation. Ann. Stat. 37, 4254–4278.
Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivariate Anal. 88, 365–411.
Lifanov, E.I., Likharev, V.A., 1983. Estimation of the covariance matrix of stationary noise. Radiotekhnika 5, 53–55.
Liu, W., Wu, W.B., 2010. Asymptotics of spectral density estimates Economet. Theor. 26, 1218–1245.
Lumley, T., Heagerty, P., 1999. Empirical adaptive variance estimators for correlated data regression. J. Roy. Stat. Soc. B 61, 459–477.
MacKinnon, J.G., White, H., 1985. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. J. Econometrics 29, 305–325.
McMurry, T.L., Politis, D.N., 2010. Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. J. Time Series Anal. 31, 471–482.
Meinshausen, N., Bühlman, P., 2006. High-dimensional graphs and variable selection with the lasso. Ann. Stat. 34, 1436–1462.
Newey, W.K., West, K.D., 1987. A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.
Ornstein, D.S., 1973. An example of a Kolmogorov automorphism that is not a Bernoulli shift. Adv. Math 10, 49–62.
Pan, J., MacKenzie, G., 2003. On modelling mean-covariance structure in longitudinal studies. Biometrika 90, 239–244.
Paul, D., 2007. Asymptotics of the leading sample eigenvalues for a spiked covariance model. Stat. Sinica. 17, 1617–1642.
Phillips, P.C.B., Solo, V., 1992. Asymptotics for linear processes. Ann. Stat. 20, 971–1001.
Pourahmadi, M., 1999. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. Biometrika 86(3), 677–690.
Pourahmadi, M., 2001. Foundations of Time Series Analysis and Prediction Theory. Wiley, New York.
Pourahmadi, M., 2011. Modeling Covariance Matrices: The GLM and Regularization Perspectives. Stat. Sci. 26(3), 369–387.

Priestley, M.B., 1981. Spectral Analysis and Time Series 1. Academic Press, London. MR0628735

Priestley, M.B., 1988. Nonlinear and Nonstationary Time Series Analysis. Academic Press, London.

Quang, A.N., 1984. On the uniqueness of the maximum-likeliwood estimate of structured covariance matrices. IEEE Trans. Acoust., Speech, Signal Processing 32(6), 1249–1251.

Rosenblatt, M., 1985. Stationary Sequences and Random Fields. Birkhäuser, Boston.

Rosenblatt, M., 2009. A comment on a conjecture of N. Wiener. Stat. Probab. Lett. 79, 347–348.

Rothman, A.J., Levina, E., Zhu, J., 2009. Generalized thresholding of large covariance matrices. J. Amer. Stat. Assoc. (Theory and Methods) 104, 177–186.

Stein, C., 1975. Estimation of a covariance matrix. In: 39th annual meeting IMS, 1975 Reitz lecture, Atlanta, Georgia.

Toeplitz, O., 1911. Zur theorie der quadratischen und bilinearan Formen von unendlichvielen, Veranderlichen. Math. Ann. 70, 351–376.

Tong, H., 1990. Non-linear Time Series: A Dynamic System Approach. Oxford University Press, Oxford.

Wachter, K.W., 1978. The strong limits of random matrix spectra for sample matrices of independent elements. Ann. Probab. 6, 1–18.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48, 817–838.

Wiener, N., 1958. Nonlinear Problems in Random Theory. MIT Press, MA.

Wu, W.B., 2005. Nonlinear system theory: Another look at dependence. Proc. Natl. Acad. Sci. USA. 102(40), 14150–14154.

Wu, W.B., 2007. M-estimation of linear models with dependent errors. Ann. Stat. 35, 495–521.

Wu, W.B., 2009. An asymptotic theory for sample covariances of Bernoulli shifts. Stochast. Proc. Appl. 119, 453–467.

Wu, W.B., 2011. Asymptotic theory for stationary processes. Stat. Interface. 4(2), 207–226.

Wu, W.B., Huang, Y., Zheng, W., 2010. Covariances estimation for long-memory processes. Adv. in Appl. Probab. 42(1), 137–157.

Wu, W.B. Min, W., 2005. On linear processes with dependent innovations. Stochast. Proc. Appl. 115(6), 939–959.

Wu, W.B., Pourahmadi, M., 2003. Nonparametric estimation of large covariance matrices of longitudinal data. Biometrika 90, 831–844.

Wu, W.B., Pourahmadi, M., 2009. Banding sample autocovariance matrices of stationary processes. Stat. Sinica 19, 1755–1768.

Xiao, H., Wu, W.B., 2011a. Asymptotic inference of autocovariances of stationary processes. preprint, available at http://arxiv.org/abs/1105.3423.

Xiao, H., Wu, W.B., 2011b. Covariance matrix estimation for stationary time series. preprint, available at http://arxiv.org/abs/1105.4563.

Yuan, M., Lin, Y., 2007. Model selection and estimation in the Gaussian graphical model. Biometrika 94, 19–35.

Zeileis, A., 2004. Econometric computing with HC and HAC covariance matrix estimators. J. Stat. Software 11(10), 117. Available from: http://www.jstatsoft.org/v11/i10/.

Zimmerman, D.L., Núñez-Antón, V., 2010. Antedependence Models for Longitudinal Data. CRC Press, New York.

This page intentionally left blank

Part IV: Time Series and Quantile Regression

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00009-0

Time Series Quantile Regressions

Zhijie Xiao
Department of Economics, Boston College, Chestnut Hill, MA 02467, USA

Abstract

Quantile information is important in time series applications. Quantile regression not only provides a method of estimating the conditional quantiles (thus the conditional distribution) of conventional time series models but also substantially expands the modeling options for time series analysis by allowing for local, quantile-specific time series dynamics. The traditional least square-based methods provide estimation for the conditional mean function. In many statistical applications, the research question is more complicate than just a few moments, and there may be valuable information about the relationship between random variables that cannot be discovered based on a simple conditional mean analysis. Quantile regression-based methods provide a complementary way to study the relationship between random variables. This chapter considers a wide range of time series quantile regression models. Quantile regressions on traditional time series models, quantile-domain local dynamic models, and time series applications are discussed.

Keywords: conditional distribution, quantile autoregression (QAR), quantile regression, time series, Value-at-Risk (VaR).

1. An introduction to quantile regression

The *quantile function* of a scalar random variable Y is the inverse of its distribution function. Like the distribution function, the quantile function provides a complete description of the statistical properties of the random variable. Similarly, the *conditional quantile function* of Y given X is the inverse of the corresponding conditional distribution function, i.e.,

$$Q_Y(\tau|X) = F_Y^{-1}(\tau|X) = \inf\{y : F_Y(y|X) \geq \tau\},$$

where $F_Y(y|X) = P(Y \leq y|X)$. The *conditional* quantile function of Y given X fully captures the relationship between Y and X.

The study of the relationship between random variables, say, X and Y, is a central issue in statistical analysis. In many applications, this is usually done by estimating some form of conditional expectation function via a Least Square (LS) regression of Y on X based on a collection of observations.

The traditional least square–based methods provide estimation for the conditional mean function. In many statistical applications, the research question is more complicate than just a few moments, there may be valuable information about the relationship between Y and X that cannot be discovered based on a simple conditional mean analysis. This problem is particularly delicate in time series, where past information may systematically affect the dynamics of the process.

Quantile regression–based methods provide an complementary way to study the relationship between X and Y. Consider the following classical linear model

$$Y_t = \theta' X_t + u_t, \quad t = 1, \ldots, n,$$

where X_t are vectors of regressors including a constant, and u_t are i.i.d. mean zero errors and are independent with X_t, a regression of the above model can be conducted based on the following optimization problem:

$$\widehat{\theta} = \min_{\theta} \sum_{t=1}^{n} \rho(Y_t - \theta' X_t), \tag{1}$$

where $\rho(\cdot)$ is a criterion (loss) function. Under appropriate regularity assumptions, solution of (1), $\widehat{\theta}$, is a consistent estimate of the vector of parameters θ^* defined as:

$$\theta^* = \min_{\theta} \mathrm{E}\rho(Y - \theta' X).$$

If we use the quadratic loss function $\rho(u) = u^2$, the ordinary LS estimator $\widehat{\theta}_{OLS}$ is obtained from (1). Solving $\theta^*_{OLS} = \min_\theta \mathrm{E}(Y - \theta' X)^2$, we have $X'\theta^*_{OLS} = \mathrm{E}(Y|X)$ – the least squares regression delivers an estimate of the conditional mean.

If we use $\rho(u) = |u|$, the Least Absolute Deviation (LAD) estimator $\widehat{\theta}_{LAD}$ is obtained. Solving $\theta^*_{LAD} = \min_\theta \mathrm{E}\left|Y - \theta' X\right|$, we have $X'\theta^*_{LAD} = \mathrm{Median}(Y|X)$ – the LAD regression delivers an estimate of the conditional median and hence is also called the median regression.

The Quantile Regression (QR) proposed by Koenker and Bassett (1978) uses an asymmetric loss function $\rho(u) = \rho_\tau(u) = u(\tau - I(u < 0))$, where $\tau \in (0, 1)$, and $I(\cdot)$ is the indicator function. Notice that $\rho_\tau(u) = (1 - \tau) I[u < 0] |u| + \tau I[u > 0] |u|$, the corresponding loss function in (1) is simply an asymmetrically weighted sum of absolute errors. Solving $\theta^*_\tau = \min_\theta \mathrm{E}\rho_\tau(Y - \theta' X)$, we obtain $X'\theta^*_\tau = Q_Y(\tau|X)$ – the (τth) quantile regression gives an estimate of the (τth) conditional quantile of Y. The criterion function $\rho_\tau(\cdot)$ is called the "check function" in the study by Koenker and Bassett (1978), and the solutions

$$\widehat{\theta}(\tau) = \min_{\theta} \sum_{t} \rho_\tau(Y_t - \theta^\top X_t) \tag{2}$$

are called the regression quantiles. Given $\widehat{\theta}(\tau)$, the τth conditional quantile function of Y_t given X_t can be estimated by

$$\hat{Q}_{Y_t}(\tau|X_t) = X_t^\top \widehat{\theta}(\tau),$$

and the conditional density of Y_t at $y = Q_{Y_t}(\tau|X_t)$ can be estimated by the difference quotients,

$$\hat{f}_{Y_t}(y|X_t) = \frac{2h}{\hat{Q}_{Y_t}(\tau + h|X_t) - \hat{Q}_{Y_t}(\tau - h|X_t)},$$

for some appropriately chosen sequence of $h = h(n) \to 0$.

Quantile regression has attracted a lot of research attention in recent years. Koenker and Hallock (2001) gave an excellent introduction of quantile regression. Also see, e.g., Cade and Noon (2003), Yu et al. (2003), and Kuan (2007) for surveys on this topic. For a systematic and complete description of quantile regression, see Koenker (2005).

This chapter focuses on time series quantile regression methods. Quantile regression not only provides a method of estimating the conditional quantiles (thus the conditional distribution) of existing time series models but also substantially expands the modeling options for time series analysis. We introduce quantile autoregressions in Section 2 and discuss quantile regressions for ARCH/GARCH models in Section 3. Quantile regressions with serially correlated residuals are considered in Section 4, and Section 5 gives a discussion on nonparametric and semiparametric time series quantile regressions. Section 6 introduces the CAViaR model and a few other dynamic quantile regression models, and Section 7 looks at extremal quantile regressions. Nonstationary time series quantile regressions are studied in Section 8. Three quantile regression applications, forecasting with quantile regressions, testing for structural changes, and portfolio construction, are briefly discussed in Section 9 to highlight the great potential of this method.

2. Quantile regression for autoregressive time series

There is a considerable literature on quantile autoregression methods including work by Weiss (1991), Knight (1989, 1998), Koul and Saleh (1995), Hercé (1996), Jureckova and Hallin (1999), and Koenker and Xiao (2004, 2006). In addition, Davis et al. (1992) and Knight (2006) studied quantile autoregression with infinite variance errors. Knight (1997) investigated second-order properties of autoregressive quantile regression estimator.

Quantile regression methods can be applied to traditional constant coefficient autoregressive models and provides estimation of the conditional quantiles in these models, and it can also be used to study new models by allowing for local, quantile-specific time series dynamics.

2.1. The classical AR model

Consider the following classical autoregressive model of order p:

$$Y_t = \theta_0 + \theta_1 Y_{t-1} + \cdots + \theta_p Y_{t-p} + u_t, \tag{3}$$

where u_t is an i.i.d. mean zero sequence with distribution function $F(\cdot)$, then the conditional distribution of Y_t (given past information) is simply a location shift of $F(\cdot)$, with conditional mean $\theta_0 + \theta_1 Y_{t-1} + \cdots + \theta_p Y_{t-p}$. Thus, the conditional quantile function of Y_t is given by

$$Q_{Y_t}(\tau|\mathcal{F}_{t-1}) = \theta_0 + \theta_1 Y_{t-1} + \cdots + \theta_p Y_{t-p} + F^{-1}(\tau),$$

where $\mathcal{F}_{t-1}$ denotes the σ-field that containing information up to time $t-1$.

Let $\theta_0(\tau) = \theta_0 + F_u^{-1}(\tau)$, $\theta(\tau) = (\theta_0(\tau), \theta_1, \ldots, \theta_p)^\top$, and $X_t = (1, Y_{t-1}, \ldots, Y_{t-p})^\top$, we may write

$$Q_{Y_t}(\tau|\mathcal{F}_{t-1}) = \theta(\tau)^\top X_t.$$

Given time series observations $\{Y_t\}_{t=1}^n$, the vector $\theta(\tau)$ can be estimated by the quantile regression (2). The asymptotic behavior of the autoregression quantiles $\widehat{\theta}(\tau)$ is summarized in the following Theorem.

THEOREM 1. *If $\{Y_t\}_{t=1}^n$ is an AR(p) process determined by (3) and $\{u_t\}$ are i.i.d. random variables with mean 0 and variance $\sigma^2 < \infty$, and the distribution function of u_t, F, has a continuous density f with $f(u) > 0$ on $\mathcal{U} = \{u : 0 < F(u) < 1\}$, then the autoregression quantiles $\widehat{\theta}(\tau)$, defined as the solution of (2), has the following limit:*

$$f[F^{-1}(\tau)]\Omega_0^{1/2}\sqrt{n}(\widehat{\theta}(\tau) - \theta(\tau)) \Rightarrow B_k(\tau),$$

where

$$\Omega_0 = E(X_t X_t^\top) = \begin{bmatrix} 1 & \mu_y' \\ \mu_y & \Omega_y \end{bmatrix},$$

$$\Omega_y = \begin{bmatrix} E(Y_t^2) & \cdots & E(Y_t Y_{t-p+1}) \\ \vdots & \ddots & \vdots \\ E(Y_t Y_{t-p+1}) & \cdots & E(Y_t^2) \end{bmatrix}, \tag{4}$$

where $\mu_y = E(Y_t) \cdot 1_{p\times 1}$, and $B_k(\tau)$ represents a k-dimensional standard Brownian Bridge, $k = p+1$.

By definition, for any fixed τ, $B_k(\tau)$ is $\mathcal{N}(0, \tau(1-\tau)I_k)$, thus

$$\sqrt{n}(\widehat{\theta}(\tau) - \theta(\tau)) \Rightarrow \mathcal{N}\left(0, \frac{\tau(1-\tau)}{f[F^{-1}(\tau)]^2}\Omega_0^{-1}\right).$$

2.2. The QAR models

In many applications, the time series dynamics can be more complicated than the classical autoregression (3), where past information (Y_{t-j}) influence only the location of the conditional distribution of Y_t. Let's look at a simple example based on a time series (Koenker, 2000; Knight, 2006) of daily temperature in Melbourne, Australia. Figure 1 is an AR(1) scatterplot of this time series. Figure 2 gives the estimated conditional density of the daily temperature conditional on the temperature of the previous day (Knight, 2006). It is quite clear from these figures that today's temperature not only affects location (and scale) of the conditional distribution of tomorrow's temperature but also the SHAPE of the conditional distribution. As the value of the conditioning variable (Y_{t-1}) increases, the conditional distribution (of Y_t) becomes bimodal!

Any attempt to diagnose or forecast series of this type requires that a mechanism be introduced to capture the empirical features of the series, or that the series be transformed in some way so that they can be analyzed by conventional models. Yet this is often much easier to say than it is to do in a satisfactory way.

We believe that quantile regression method can be used to address some of these problems. An important extension of the classical constant coefficient time series model is the Quantile Autoregression (QAR) model (Koenker and Xiao, 2006). Given a time series $\{Y_t\}$, let $\mathcal{F}_t$ be the σ-field generated by $\{Y_s, s \leq t\}$, $\{Y_t\}$ is a pth order QAR process if

$$Q_{Y_t}(\tau|\mathcal{F}_{t-1}) = \theta_0(\tau) + \theta_1(\tau)Y_{t-1} + \cdots + \theta_p(\tau)Y_{t-p}. \tag{5}$$

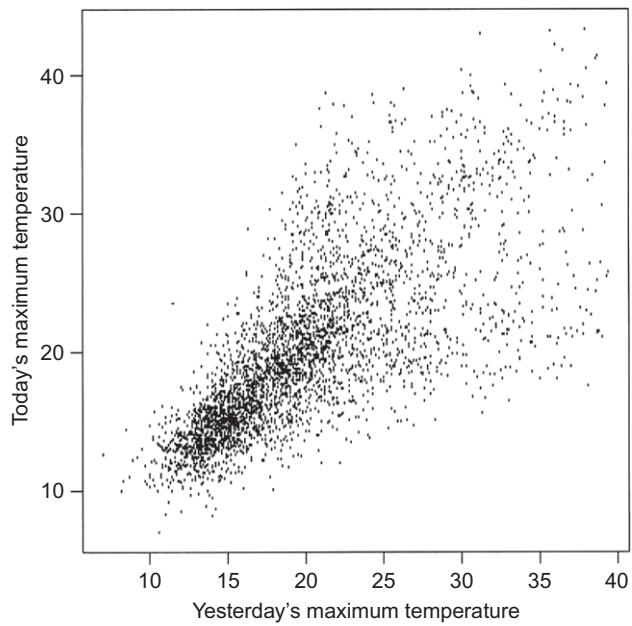

Fig. 1. Scatterplot of Melbourne temperature.

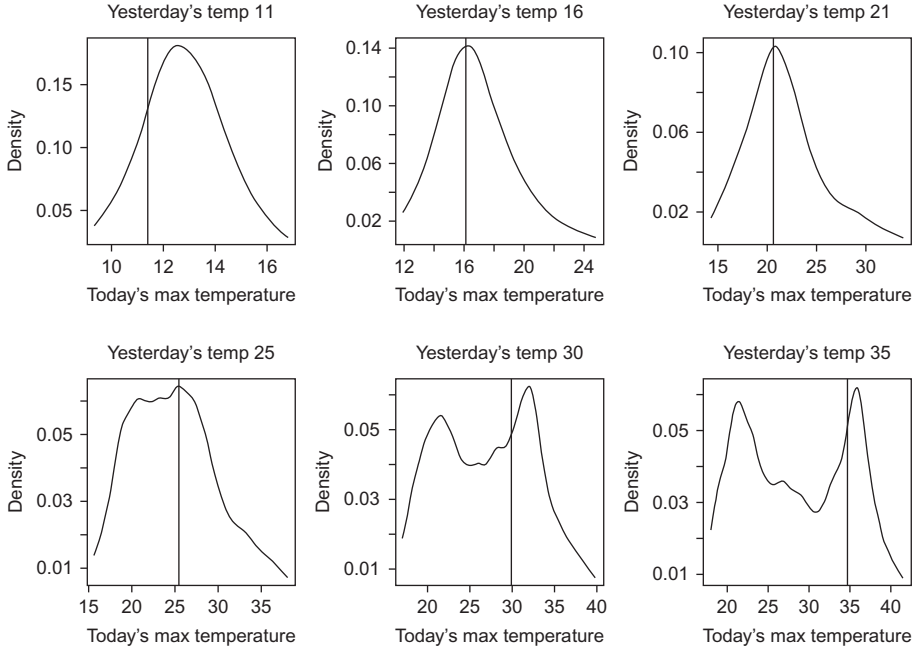

Fig. 2. Estimated conditional densities (Koenker, 2006).

This implies, of course, that the right hand side of (5) is monotonically increasing in τ. In the above QAR model, the autoregressive coefficients may be τ-dependent and thus can vary over different quantiles of the conditional distribution. Consequently, the conditioning variables not only shift the location of the distribution of Y_t but also may alter the scale and shape of the conditional distribution. The QAR models play a useful role in expanding the modeling territory of the classical autoregressive time series models, and the classical AR(p) model can be viewed as a special case of QAR by setting $\theta_j(\tau)$ ($j = 1, \ldots, p$) to constants.

The formulation in (5) reveals that the QAR model may be interpreted as a somewhat special form of random coefficient autoregressive (RCAR) model:

$$Y_t = \theta_0(U_t) + \theta_1(U_t)Y_{t-1} + \cdots + \theta_p(U_t)Y_{t-p}, \tag{6}$$

where $\{U_t\}$ is a sequence of i.i.d. standard uniform random variables. In contrast to most of the literature on RCAR models, in which the coefficients are typically assumed to be stochastically independent of one another, the QAR model has coefficients that are functionally dependent.

To illustrate some important features of the QAR process, we may consider the following simple QAR (1) process,

$$Y_t = \alpha_t Y_{t-1} + u_t, \tag{7}$$

where $u_t = \theta_0(U_t)$ with $\theta_0(U_t) = F^{-1}(U_t)$, $F(\cdot)$ is some distribution function, and

$$\alpha_t = \begin{cases} \frac{1}{2} + U_t, & U_t < \frac{1}{2}, \\ 1, & U_t \geq \frac{1}{2}. \end{cases}$$

In this model, if $U_t \geq 1/2$, the model generates the Y_t according to a unit root model, but for smaller realizations of the innovation, we have a mean reversion tendency. Thus, the model exhibits a form of asymmetric persistence in the sense that sequences of strongly positive innovations tend to reinforce its unit root like behavior, whereas occasional negative realizations induce mean reversion and thus undermine the persistence of the process. In fact, Y_t is covariance stationary and satisfies a central limit theorem. Thus, a quantile autoregressive process may allow for some transient forms of explosive behavior while maintaining stationarity in the long run.

Denote $X_t = (1, Y_{t-1}, \ldots, Y_{t-p})^\top$, and $\theta(\tau) = \left(\theta_0(\tau), \theta_1(\tau), \ldots, \theta_p(\tau)\right)^\top$, the quantile autoregressive model (5) can be estimated by the conventional quantile regression technique through (2).

To facilitate the asymptotic analysis, we reformulate the QAR(p) model (6) in the more conventional random coefficient notation as,

$$Y_t = \mu_0 + \alpha_{1,t} Y_{t-1} + \cdots + \alpha_{p,t} Y_{t-p} + u_t, \tag{8}$$

where $\mu_0 = E\theta_0(U_t)$, $u_t = \theta_0(U_t) - \mu_0$, and $\alpha_{j,t} = \theta_j(U_t)$, for $j = 1, \ldots, p$. Thus, $\{u_t\}$ is an i.i.d. sequence of random variables with distribution function $F(\cdot) = \theta_0^{-1}(\cdot + \mu_0)$, and the $\alpha_{j,t}$ coefficients are functions of this u_t innovation random variable. The QAR(p) process (8) can be expressed as an p-dimensional vector autoregression process of order 1:

$$\mathbf{Y}_t = \Gamma + A_t \mathbf{Y}_{t-1} + \mathbf{V}_t$$

with

$$\Gamma = \begin{bmatrix} \mu_0 \\ 0_{p-1} \end{bmatrix}, \quad A_t = \begin{bmatrix} A_{p-1,t} & \alpha_{p,t} \\ I_{p-1} & 0_{p-1} \end{bmatrix}, \quad \mathbf{V}_t = \begin{bmatrix} u_t \\ 0_{p-1} \end{bmatrix},$$

where $A_{p-1,t} = [\alpha_{1,t}, \ldots, \alpha_{p-1,t}], \mathbf{Y}_t = [Y_t, \ldots, Y_{t-p+1}]^\top$, and 0_{p-1} is the $(p-1)$-dimensional vector of zeros. Koenker and Xiao (2006) studied the QAR model under the following conditions:

A.1 $\{u_t\}$ are i.i.d. random variables with mean 0 and variance $\sigma^2 < \infty$. The distribution function of u_t, F, has a continuous density f with $f(u) > 0$ on $\mathcal{U} = \{u : 0 < F(u) < 1\}$.

A.2 Let $E(A_t \otimes A_t) = \Omega_A$, the eigenvalues of Ω_A have moduli less than unity.

A.3 Denote the conditional distribution function $\Pr[y_t < \cdot | \mathcal{F}_{t-1}]$ as $F_{t-1}(\cdot)$ and its derivative as $f_{t-1}(\cdot)$, f_{t-1} is uniformly integrable on $\mathcal{U}$.

THEOREM 2. *Under assumptions A.1–A.3, (1) the QAR(p) process Y_t given by (8) is covariance stationary and satisfies a central limit theorem*

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} (Y_t - \mu_y) \Rightarrow N\left(0, \omega_y^2\right),$$

where $\mu_y = \mu_0 / \left(1 - \sum_{j=1}^{p} \mu_j\right)$, $\omega_y^2 = \lim n^{-1} E[\sum_{t=1}^{n} (y_t - \mu_y)]^2$, *and* $\mu_j = E(\alpha_{j,t})$, $j = 1, \ldots, p$. *(2) The autoregression quantile process* $\widehat{\theta}(\tau)$ *has the following limiting representation:*

$$\Sigma^{-1/2} \sqrt{n}(\widehat{\theta}(\tau) - \theta(\tau)) \Rightarrow B_k(\tau),$$

where $\Sigma = \Omega_1^{-1} \Omega_0 \Omega_1^{-1}$, $\Omega_1 = \lim n^{-1} \sum_{t=1}^{n} f_{t-1}[F_{t-1}^{-1}(\tau)] X_t X_t^{\top}$, $\Omega_0 = E\left(X_t X_t^{\top}\right)$, *and* $B_k(\tau)$ *represents a k-dimensional standard Brownian Bridge,* $k = p + 1$.

From Theorem 2, we have, for fixed τ, the limiting distribution of the QAR estimator given by

$$\sqrt{n}(\widehat{\theta}(\tau) - \theta(\tau)) \Rightarrow \mathcal{N}\left(0, \tau(1-\tau)\Omega_1^{-1} \Omega_0 \Omega_1^{-1}\right).$$

The QAR models expand the modeling options for time series that display asymmetric dynamics and allows for local persistency. The models can capture systematic influences of conditioning variables on the location, scale, and shape of the conditional distribution of the response and, therefore, constitute a significant extension of classical constant coefficient linear time series models.

Quantile varying coefficients indicate the existence of conditional heteroskedasticity. Given the QAR process (6), let $\theta_0 = \mathrm{E}[\theta_0(U_t)]$, $\theta_1 = \mathrm{E}[\theta_1(U_t)], \ldots, \theta_p = \mathrm{E}\left[\theta_p(U_t)\right]$, and

$$V_t = \theta_0(U_t) - \mathrm{E}\theta_0(U_t) + [\theta_1(U_t) - \mathrm{E}\theta_1(U_t)]\, Y_{t-1} + \cdots + \left[\theta_p(U_t) - \mathrm{E}\theta_p(U_t)\right] Y_{t-p}.$$

The QAR process can be rewritten as

$$Y_t = \theta_0 + \theta_1 Y_{t-1} + \cdots + \theta_p Y_{t-p} + V_t, \tag{9}$$

where V_t is martingale difference sequence. The QAR process is a weak-sense AR process with conditional heteroskedasticity.

What is the difference between a QAR process and an AR process with ARCH (or GARCH) errors? In short, the ARCH-type model focuses only on the first two moments, whereas the QAR model goes beyond the second moment and allows for more flexible structure in higher moments. Both models allow for conditional heterokedasticity and they are similar in the first two moments, but they can be quite different beyond conditional variance.

The classical time series analysis based on autocorrelations (and partial autocorrelations, etc.) only requires that the residuals are uncorrelated (martingale difference sequence). As we show in (9), the autocovariance structure of the QAR process (6) is the same as that of a fixed coefficient AR(p) process. Thus, if we consider two different QAR(p) processes:

$$Y_{1,t} = \theta_{10}(U_t) + \theta_{11}(U_t) Y_{1,t-1} + \cdots + \theta_{1p}(U_t) Y_{1,t-p}$$

and

$$Y_{2,t} = \theta_{20}(U_t) + \theta_{21}(U_t) Y_{2,t-1} + \cdots + \theta_{2p}(U_t) Y_{2,t-p},$$

then if $\mathrm{E}\big[\theta_{1j}(U_t)\big] = \mathrm{E}\big[\theta_{2j}(U_t)\big]$, their autocorrelation structures are the same. Consequently, the classical time series analysis technique will identify QAR processes with different **dependence** structures as the same (fixed coefficient) AR(p) process. In this case, the QAR technique helps to reveal valuable additional information that the classical time series analysis may ignore. The QAR method provides a very useful complement to the classical analysis in identifying time series with different local behavior. See Knight (2006) for related discussions.

A simple high-level assumption that we made on the QAR process is monotonicity of the right-hand side of (5). The monotonicity of the conditional quantile functions imposes some discipline on the forms taken by the θ functions. It usually imposes restrictions on the domain of the random variable Y_t unless Y_t is a traditional constant coefficient process. It requires that the domain of the random variables (or appropriately transformed versions) are bounded at least in one direction (say, non-negative).

If the monotonicity assumption does not hold, the results in Theorem 2 need to be modified. We may still consider the linear quantile regression, but treating $X_t'\widehat{\theta}(\tau)$ as an approximation for $Q_{Y_t}(\tau|\mathcal{F}_{t-1})$. In this case, $\widehat{\theta}(\tau)$ will converge to some pseudo-parameter $\overline{\theta}(\tau)$ that minimizes some distance between $X_t'\theta$ and $Q_{Y_t}(\tau|\mathcal{F}_{t-1})$, i.e.,

$$\widehat{\theta}(\tau) \to_p \overline{\theta}(\tau) = \arg\min_{\theta} \mathrm{E}d\left(X_t'\theta, Q_{Y_t}(\tau|\mathcal{F}_{t-1})\right),$$

where the distance is defined as $d(X_t'\theta, Q_{Y_t}(\tau|\mathcal{F}_{t-1})) = \mathrm{E}\{(\delta - |\varepsilon_{t\tau}|)1(|\varepsilon_{t\tau}| < \delta) |\mathcal{F}_{t-1}\}$, with $\delta(\theta, X_t) = |X_t'\theta - Q_{Y_t}(\tau|\mathcal{F}_{t-1})|$, and $\varepsilon_{t\tau} = Y_t - Q_{Y_t}(\tau|\mathcal{F}_{t-1})$.

One can establish asymptotic normality of $\widehat{\theta}(\tau)$ around $\overline{\theta}(\tau)$. This is similar to the general theory of AR(p) estimation under misspecification. The estimated linear QAR model serves as a local approximation device for the global model. Statistical inference can still be conducted, but the limiting distribution needs to be modified to accommodate the possible misspecification. In particular, without monotonicity assumption, under regularity assumptions, the following asymptotic representation (and thus asymptotic normality) can be obtained:

$$\sqrt{n}\left(\widehat{\theta}(\tau) - \overline{\theta}(\tau)\right) = V_n(\tau)^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n} X_t\psi_\tau(u_{t\tau}^*) + o_p(1),$$

where $V_n(\tau) = n^{-1}\sum_{t=1}^{n} f_t\left(X_t'\overline{\theta}(\tau)\right)X_tX_t^\top$, and $u_{t\tau}^* = y_t - X_t'\overline{\theta}(\tau)$, $\psi_\tau(u) = \tau - I(u < 0)$, extending the result of Angrist et al. (2005) from i.i.d. case to time series models. Simulation-based methods such as subsampling may be used to conduct statistical inference for the QAR models under misspecification.

Despite the possible crossing of quantile curves, the linear QAR model provides a convenient and useful local approximation to global nonlinear QAR models. Such simplified QAR models can still deliver important insight about dynamics, e.g., adjustment asymmetries, in time series observations and thus provide a useful tool in empirical diagnostic time series analysis. See Koenker and Xiao (2006) and discussions on QAR in the issue of JASA (Vol. 101, 2006) for more details.

2.3. Nonlinear QAR models

More complicated functional forms with nonlinearity can be considered for the conditional quantile function if we are interested in the global behavior of the time series. The absence of monotonicity implies that a more complicated functional form with nonlinearity is needed for $Q_{Y_t}(\tau|X_t)$. If the τth conditional quantile function of Y_t is given by

$$Q_{Y_t}(\tau|\mathcal{F}_{t-1}) = H(X_t;\theta(\tau)),$$

where X_t is the vector containing lagged Ys, we may estimate the vector of parameters $\theta(\tau)$ (and thus the conditional quantile of Y_t) by the following nonlinear quantile regression:

$$\min_{\theta} \sum_t \rho_\tau(Y_t - H(X_t,\theta)). \tag{10}$$

Let $\varepsilon_{t\tau} = y_t - H(x_t,\theta(\tau))$, $\dot{H}_\theta(x_t,\theta) = \partial H(x_t;\theta)/\partial\theta$, we assume that:

$$V_n(\tau) = \frac{1}{n}\sum_t f_t(Q_{Y_t}(\tau|X_t))\dot{H}_\theta(X_t,\theta(\tau))\dot{H}_\theta(X_t,\theta(\tau))^\top \xrightarrow{P} V(\tau),$$

$$\Omega_n(\tau) = \frac{1}{n}\sum_t \dot{H}_\theta(X_t,\theta(\tau))\dot{H}_\theta(X_t,\theta(\tau))^\top \xrightarrow{P} \Omega(\tau),$$

and

$$\frac{1}{\sqrt{n}}\sum_t \dot{H}_\theta(x_t,\theta(\tau))\psi_\tau(\varepsilon_{t\tau}) \Rightarrow N(0,\tau(1-\tau)\Omega(\tau)),$$

where $V(\tau)$ and $\Omega(\tau)$ are nonsingular, then under appropriate assumptions, the nonlinear QAR estimator $\widehat{\theta}(\tau)$ defined as solution of (10) is root-n consistent and

$$\sqrt{n}\left(\widehat{\theta}(\tau) - \theta(\tau)\right) \Rightarrow N(0,\tau(1-\tau)V(\tau)^{-1}\Omega(\tau)V(\tau)^{-1}). \tag{11}$$

In practice, one may employ parametric copula models to generate nonlinear-in-parameters QAR models (see, e.g., Bouyé and Salmon (2008) and Chen et al. (2009)). Copula-based Markov models provide a rich source of potential nonlinear dynamics describing temporal dependence and tail dependence. If we consider, for example, a first-order strictly stationary Markov process, $\{Y_t\}_{t=1}^n$, whose probabilistic properties are determined by the joint distribution of Y_{t-1} and Y_t, say, $G^*(y_{t-1},y_t)$, and suppose that $G^*(y_{t-1},y_t)$ has continuous marginal distribution function $F^*(\cdot)$, then by Sklar's Theorem, there exists an unique copula function $C^*(\cdot,\cdot)$, such that

$$G^*(y_{t-1},y_t) \equiv C^*(F^*(y_{t-1}),F^*(y_t)),$$

where the copula function $C^*(\cdot,\cdot)$ is a bivariate probability distribution function with uniform marginals. Differentiating $C^*(u,v)$ with respect to u, and evaluate at $u = F^*(x), v = F^*(y)$, we obtain the conditional distribution of Y_t given $Y_{t-1} = x$:

$$\Pr[Y_t < y | Y_{t-1} = x] = \left.\frac{\partial C^*(u,v)}{\partial u}\right|_{u=F^*(x),v=F^*(y)} \equiv C_1^*(F^*(x), F^*(y)).$$

For any $\tau \in (0,1)$, solving $\tau = \Pr[Y_t < y | Y_{t-1} = x] \equiv C_1^*(F^*(x), F^*(y))$ for y (in terms of τ), we obtain the τth conditional quantile function of Y_t given $Y_{t-1} = x$:

$$Q_{Y_t}(\tau|x) = F^{*-1}(C_1^{*-1}(\tau; F^*(x))),$$

where $F^{*-1}(\cdot)$ signifies the inverse of $F^*(\cdot)$ and $C_1^{*-1}(\cdot;u)$ is the partial inverse of $C_1^*(u,v)$ with respect to $v = F^*(y_t)$.

In practice, neither the true copula function $C^*(\cdot,\cdot)$ nor the true marginal distribution function $F^*(\cdot)$ of $\{Y_t\}$ is known. If we model both parametrically by $C(\cdot,\cdot;\alpha)$ and $F(y;\beta)$, then the τth conditional quantile function of Y_t, $Q_{Y_t}(\tau|x)$ becomes a function of the unknown parameters α and β, i.e.,

$$Q_{Y_t}(\tau|x) = F^{-1}(C_1^{-1}(\tau; F(x,\beta),\alpha),\beta).$$

Denoting $\theta = (\alpha', \beta')'$ and $h(x,\alpha,\beta) \equiv C_1^{-1}(\tau; F(x,\beta),\alpha)$, we will write,

$$Q_{Y_t}(\tau|x) = F^{-1}(h(x,\alpha,\beta),\beta) \equiv H(x;\theta). \tag{12}$$

For example, if we consider the Clayton copula:

$$C(u,v;\alpha) = [u^{-\alpha} + v^{-\alpha} - 1]^{-1/\alpha}, \quad \text{where} \quad \alpha > 0.$$

one can easily verify that the τth conditional quantile function of U_t given u_{t-1} is

$$Q_{U_t}(\tau|u_{t-1}) = [(\tau^{-\alpha/(1+\alpha)} - 1)u_{t-1}^{-\alpha} + 1]^{-1/\alpha}$$

See Bouyé and Salmon (2008) for additional examples of copula-based conditional quantile functions.

Although the quantile function specification in the above representation assumes the parameters to be identical across quantiles, we may permit the estimated parameters to vary with τ and thus extending the original copula-based QAR models to capture a wide range of systematic influences of conditioning variables on the conditional distribution of the response. By varying the choice of the copula specification, we can induce a wide variety of nonlinear QAR dependence, and the choice of the marginal enables us to consider a wide range of possible tail behavior as well. In many financial time series applications, the nature of the temporal dependence varies over the quantiles of the conditional distribution. Chen et al. (2009) studied the asymptotic properties of the copula-based nonlinear quantile autoregression.

REMARK 1. We could even further relax the assumption of the conditional quantile function and allow for a nonparametric specification. See Section 5 on discussions of nonparametric QR. □

REMARK 2. ARMA models may also be analyzed via nonlinear QR in a similar way. □

3. Quantile regression for ARCH and GARCH models

ARCH and GARCH models have proven to be highly successful in modeling financial data. Estimators of volatilities and quantiles based on ARCH and GARCH models are now widely used in finance applications. Koenker and Zhao (1996) studied quantile regression for linear ARCH models. They consider the following linear ARCH(p) process

$$u_t = \sigma_t \cdot \varepsilon_t, \sigma_t = \gamma_0 + \gamma_1 |u_{t-1}| + \cdots + \gamma_p |u_{t-p}|, \tag{13}$$

where $0 < \gamma_0 < \infty$, $\gamma_1, \ldots, \gamma_p \geq 0$, and ε_t are i.i.d. (0,1) random variables with pdf $f(\cdot)$ and CDF $F(\cdot)$. Let $Z_t = (1, |u_{t-1}|, \ldots, |u_{t-q}|)^\top$ and $\gamma(\tau) = (\gamma_0 F^{-1}(\tau), \gamma_1 F^{-1}(\tau), \ldots, \gamma_q F^{-1}(\tau))^\top$, the conditional quantiles of u_t is given by

$$Q_{u_t}(\tau|\mathcal{F}_{t-1}) = \gamma_0(\tau) + \gamma_1(\tau)|u_{t-1}| + \cdots + \gamma_p(\tau)|u_{t-p}| = \gamma(\tau)^\top Z_t$$

and can be estimated by the following linear quantile regression of u_t on Z_t:

$$\min_{\gamma} \sum_t \rho_\tau(u_t - \gamma^\top Z_t), \tag{14}$$

where $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_q)^\top$. The asymptotic behavior of the above quantile regression estimator is given in the following theorem (Koenker and Zhao, 1996).

THEOREM 3. *Suppose that u_t is given by model (13), f is bounded and continuous, $f(F^{-1}(\tau)) > 0$ for any $0 < \tau < 1$. In addition, $E|u_t|^{2+\delta} < \infty$, then the regression quantiles $\widehat{\gamma}(\tau)$ of (14) has the following Bahadur representation*

$$\sqrt{n}(\widehat{\gamma}(\tau) - \gamma(\tau)) = \frac{\Sigma_1^{-1}}{f(F^{-1}(\tau))} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} Z_t \psi_\tau(\varepsilon_{t\tau}) + o_p(1),$$

where $\Sigma_1 = EZ_t Z_t'/\sigma_t$ and $\varepsilon_{t\tau} = \varepsilon_t - F^{-1}(\tau)$. Consequently,

$$\sqrt{n}(\widehat{\gamma}(\tau) - \gamma(\tau)) = N\left(0, \frac{\tau(1-\tau)}{f(F^{-1}(\tau))^2} \Sigma_1^{-1} \Sigma_0 \Sigma_1^{-1}\right), \quad \text{with } \Sigma_0 = EZ_t Z_t'.$$

In many applications, conditional heteroskedasticity is modeled on the residuals of a regression. For example, we may consider the following AR–ARCH model:

$$Y_t = \alpha' X_t + u_t, \tag{15}$$

where $X_t = (1, Y_{t-1}, \ldots, Y_{t-p})^\top$, $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_p)^\top$, and u_t is a linear ARCH(p) process given by model (13). The conditional quantiles of Y_t is then given by

$$Q_{Y_t}(\tau|\mathcal{F}_{t-1}) = \alpha' X_t + \gamma(\tau)^\top Z_t. \tag{16}$$

One way to estimate the above model is to construct a *joint* estimation of α and $\gamma(\tau)$ based on *nonlinear* quantile regression. Alternatively, we may consider a two-step procedure that estimates α in the first step and then estimates $\gamma(\tau)$ based on the estimated residuals. The two-step procedure is usually less efficient because the preliminary estimation of α may affect the second-step estimation of $\gamma(\tau)$, but it is computationally much simpler and is widely used in empirical applications. Koenker and Zhao (1996) studied the two-step estimation, and Theorem 4 summarizes the results.

THEOREM 4. *Suppose that Y_t is given by (15) and (13), and conditions of Theorem 3 holds, $\widehat{\alpha}$ is a root-n consistent estimator, and*

$$\widetilde{\gamma}(\tau) = \arg\min_{\gamma} \sum_t \rho_\tau(\widehat{u}_t - \gamma^\top \widehat{Z}_t),$$

where $\widehat{Z}_t = (1, |\widehat{u}_{t-1}|, \ldots, |\widehat{u}_{t-q}|)^\top$, $\widehat{u}_t = Y_t - \widehat{\alpha}' X_t$, then

$$\sqrt{n}\,(\widetilde{\gamma}(\tau) - \gamma(\tau)) = \frac{\Sigma_1^{-1}}{f(F^{-1}(\tau))} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} Z_t \psi_\tau(\varepsilon_{t\tau}) + \Sigma_1^{-1} G_1 \sqrt{n}\,(\widehat{\alpha} - \alpha) + o_p(1)$$

with $G_1 = E\big(\sigma_t^{-1} Z_t (X_t - B_t \gamma(\tau))^\top\big)$, and $B_t = (0, sign(u_{t-1}) X_{t-1}, \ldots, sign(u_{t-p}) X_{t-p})$. If f is symmetric about zero, and $\alpha_0 = 0$, then $G_1 = 0$, and thus

$$\sqrt{n}\,(\widetilde{\gamma}(\tau) - \gamma(\tau)) = N\left(0, \frac{\tau(1-\tau)}{f(F^{-1}(\tau))^2} \Sigma_1^{-1} \Sigma_0 \Sigma_1^{-1}\right).$$

ARCH models are easier to estimate, but cannot parsimoniously capture the persistent influence of long past shocks comparing to the GARCH models. However, quantile regression GARCH models are highly nonlinear and thus complicated to estimate. In particular, the quantile estimation problem in GARCH models corresponds to a restricted nonlinear quantile regression and conventional nonlinear quantile regression techniques are not directly applicable.

Xiao and Koenker (2009) studied quantile regression estimation of the following linear GARCH(p, q) model:

$$u_t = \sigma_t \cdot \varepsilon_t, \tag{17}$$

$$\sigma_t = \beta_0 + \beta_1 \sigma_{t-1} + \cdots + \beta_p \sigma_{t-p} + \gamma_1 |u_{t-1}| + \cdots + \gamma_q \left|u_{t-q}\right|. \tag{18}$$

Let $\mathcal{F}_{t-1}$ represents information up to time $t-1$, the τ th conditional quantile of u_t is given by

$$Q_{u_t}(\tau|\mathcal{F}_{t-1}) = \theta(\tau)^\top Z_t, \tag{19}$$

where $Z_t = (1, \sigma_{t-1}, \ldots, \sigma_{t-p}, |u_{t-1}|, \ldots, |u_{t-q}|)^\top$ and $\theta(\tau)^\top = (\beta_0, \beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_q)F^{-1}(\tau)$.

As Z_t contains σ_{t-k} $(k = 1, \ldots, p)$, which in turn depends on unknown parameters $\theta = (\beta_0, \beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_q)$, we may write Z_t as $Z_t(\theta)$ to emphasize the nonlinearity and its dependence on θ. If we use the following nonlinear quantile regression

$$\min_{\theta} \sum_t \rho_\tau(u_t - \theta^\top Z_t(\theta)), \tag{20}$$

for a fixed τ in isolation, consistent estimate of θ cannot be obtained because it ignores the global dependence of the σ_{t-k}'s on the entire function $\theta(\cdot)$. If the dependence structure of u_t is characterized by (17) and (18), we can consider the following restricted quantile regression instead of (20):

$$\left(\widehat{\pi}, \widehat{\theta}\right) = \begin{cases} \arg\min_{\pi,\theta} \sum_i \sum_t \rho_{\tau_i}(u_t - \pi_i^\top Z_t(\theta)) \\ s.t.\, \pi_i = \theta(\tau_i) = \theta F^{-1}(\tau_i). \end{cases}$$

Estimation of this global restricted nonlinear quantile regression is complicated. Xiao and Koenker (2009) propose a simpler two-stage estimator that both incorporates the global restrictions and also focuses on the local approximation around the specified quantile. The proposed estimation consists of the following two steps: (i) The first step considers a global estimation to incorporate the global dependence of the latent σ_{t-k}'s on θ. (ii) Then, using results from the first step, we focus on the specified quantile to find the best local estimate for the conditional quantile. Let

$$A(L) = 1 - \beta_1 L - \cdots - \beta_p L^p, \quad B(L) = \gamma_1 + \cdots + \gamma_q L^{q-1},$$

under regularity assumptions ensuring that $A(L)$ is invertible, we obtain an ARCH(∞) representation for σ_t:

$$\sigma_t = a_0 + \sum_{j=1}^{\infty} a_j \left|u_{t-j}\right|. \tag{21}$$

For identification, we normalize $a_0 = 1$. Substituting the above ARCH(∞) representation into (17) and (18), we have

$$u_t = \left(a_0 + \sum_{j=1}^{\infty} a_j \left|u_{t-j}\right|\right) \varepsilon_t \tag{22}$$

and

$$Q_{u_t}(\tau|\mathcal{F}_{t-1}) = \alpha_0(\tau) + \sum_{j=1}^{\infty} \alpha_j(\tau) \left|u_{t-j}\right|,$$

where $\alpha_j(\tau) = a_j Q_{\varepsilon_t}(\tau)$, $j = 0, 1, 2, \ldots$.

Let $m = m(n)$ be a truncation parameter, we may consider the following truncated quantile autoregression:

$$Q_{u_t}(\tau|\mathcal{F}_{t-1}) \approx a_0(\tau) + a_1(\tau)\,|u_{t-1}| + \cdots + a_m(\tau)\,|u_{t-m}|\,.$$

By choosing m suitably, small relative to the sample size n, but large enough to avoid serious bias, we obtain a sieve approximation for the GARCH model.

One could estimate the conditional quantiles simply using a sieve approximation:

$$\check{Q}_{u_t}(\tau|\mathcal{F}_{t-1}) = \widehat{a}_0(\tau) + \widehat{a}_1(\tau)\,|u_{t-1}| + \cdots + \widehat{a}_m(\tau)\,|u_{t-m}|\,,$$

where $\widehat{a}_j(\tau)$ are the quantile autoregression estimates. Under regularity assumptions,

$$\check{Q}_{u_t}(\tau|\mathcal{F}_{t-1}) = Q_{u_t}(\tau|\mathcal{F}_{t-1}) + O_p(m/\sqrt{n}).$$

However, Monte Carlo evidence indicates that the simple sieve approximation does not directly provide a good estimator for the GARCH model, but it serves as an adequate preliminary estimator. Because the first step of estimation focuses on the global model, it is desirable to use information over multiple quantiles in estimation. Combining information over multiple quantiles helps us to obtain globally coherent estimate of the scale parameters.

Suppose that we estimate the mth-order quantile autoregression

$$\widetilde{\alpha}(\tau) = \arg\min_{\alpha} \sum_{t=m+1}^{n} \rho_\tau \left(u_t - \alpha_0 - \sum_{j=1}^{m} \alpha_j \left|u_{t-j}\right| \right) \tag{23}$$

at quantiles $(\tau_1, \ldots, \tau_K)$, and obtain estimates $\widetilde{\alpha}(\tau_k)$, $k = 1, \ldots, K$. Let $\widetilde{a}_0 = 1$ in accordance with the identification assumption. Denote

$$\mathbf{a} = [a_1, \ldots, a_m, q_1, \ldots, q_K]^\top, \quad \widetilde{\boldsymbol{\pi}} = \left[\widetilde{\alpha}(\tau_1)^\top, \ldots, \widetilde{\alpha}(\tau_K)^\top\right]^\top,$$

where $q_k = Q_{\varepsilon_t}(\tau_k)$, and

$$\phi(\mathbf{a}) = g \otimes \alpha = [q_1, a_1 q_1, \ldots, a_m q_1, \ldots,\ q_K, a_1 q_K, \ldots, a_m q_K]^\top,$$

where $g = [q_1, \ldots, q_K]^\top$ and $\alpha = [1, a_1, a_2, \ldots, a_m]^\top$, we consider the following estimator for the vector $\mathbf{a}$ that combines information over the K quantile estimates based on the restrictions $\alpha_j(\tau) = a_j Q_{\varepsilon_t}(\tau)$:

$$\widetilde{\mathbf{a}} = \arg\min_{\mathbf{a}} \left(\widetilde{\boldsymbol{\pi}} - \phi(\mathbf{a})\right)^\top A_n \left(\widetilde{\boldsymbol{\pi}} - \phi(\mathbf{a})\right), \tag{24}$$

where A_n is a $(K(m+1)) \times (K(m+1))$ positive definite matrix. Denoting $\widetilde{\mathbf{a}} = (\widetilde{a}_0, \ldots, \widetilde{a}_m)$, σ_t can be estimated by

$$\widetilde{\sigma}_t = \widetilde{a}_0 + \sum_{j=1}^{m} \widetilde{a}_j \left|u_{t-j}\right|.$$

In the second step, we perform a quantile regression of u_t on $\widetilde{Z}_t = (1, \widetilde{\sigma}_{t-1}, \ldots, \widetilde{\sigma}_{t-p}, |u_{t-1}|, \ldots, |u_{t-q}|)^\top$ by

$$\min_{\theta} \sum_t \rho_\tau(u_t - \theta^\top \widetilde{Z}_t), \tag{25}$$

the two-step estimator of $\theta(\tau)^\top = (\beta_0(\tau), \beta_1(\tau), \ldots, \beta_p(\tau), \gamma_1(\tau), \ldots, \gamma_q(\tau))$ is then given by solution of (25), $\widehat{\theta}(\tau)$, and the τth conditional quantile of u_t can be estimated by

$$\widehat{Q}_{u_t}(\tau|\mathcal{F}_{t-1}) = \widehat{\theta}(\tau)^\top \widetilde{Z}_t.$$

Iteration can be applied to the above procedure for further improvement.

Let $\widetilde{\alpha}(\tau)$ be the solution of (23), then under appropriate Assumptions, we have

$$\|\widetilde{\alpha}(\tau) - \alpha(\tau)\|^2 = O_p(m/n). \tag{26}$$

and for any $\lambda \in \mathcal{R}^{m+1}$,

$$\frac{\sqrt{n}\lambda^\top(\widetilde{\alpha}(\tau) - \alpha(\tau))}{\sigma_\lambda} \Rightarrow N(0,1),$$

where $\sigma_\lambda^2 = f_\varepsilon\left(F_\varepsilon^{-1}(\tau)\right)^{-2} \lambda^\top D_n^{-1} \Sigma_n(\tau) D_n^{-1} \lambda$, and

$$D_n = \left[\frac{1}{n}\sum_{t=m+1}^{n} \frac{x_t x_t^\top}{\sigma_t}\right], \quad \Sigma_n(\tau) = \frac{1}{n}\sum_{t=m+1}^{n} x_t x_t^\top \psi_\tau^2(u_{t\tau}),$$

where $x_t = (1, |u_{t-1}|, \ldots, |u_{t-m}|)^\top$.

Define

$$G = \left.\frac{\partial \phi(\mathbf{a})}{\partial \mathbf{a}^\top}\right|_{\mathbf{a}=\mathbf{a}_0} = \dot{\phi}(\mathbf{a}_0) = \left[g_0 \otimes J_m \vdots I_K \otimes \alpha_0\right], \quad g_0 = \begin{bmatrix} Q_{\varepsilon_t}(\tau_1) \\ \cdots \\ Q_{\varepsilon_t}(\tau_K) \end{bmatrix},$$

where g_0 and α_0 are the true values of vectors $g = [q_1, \cdots, q_K]^\top$ and $\alpha = [1, a_1, a_2, \ldots, a_m]^\top$, and

$$J_m = \begin{bmatrix} 0 & \cdots & 0 \\ 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$$

is an $(m+1) \times m$ matrix and I_K is an K-dimensional identity matrix, under regularity assumptions, the minimum distance estimator $\widetilde{\mathbf{a}}$ solving (24) has the following asymptotic representation:

$$\sqrt{n}(\widehat{\mathbf{a}} - \mathbf{a}_0) = \left[G^\top A_n G\right]^{-1} G^\top A_n \sqrt{n}\left(\widetilde{\boldsymbol{\pi}} - \boldsymbol{\pi}\right) + o_p(1),$$

where

$$\sqrt{n}(\widetilde{\boldsymbol{\pi}} - \boldsymbol{\pi}) = -\frac{1}{\sqrt{n}} \sum_{t=m+1}^{n} \begin{bmatrix} \left(D_n^{-1} x_t \dfrac{\psi_{\tau_1}(u_{t\tau_1})}{f_\varepsilon\left(F_\varepsilon^{-1}(\tau_1)\right)} \right) \\ \cdots \\ \left(D_n^{-1} x_t \dfrac{\psi_{\tau_k}(u_{t\tau_k})}{f_\varepsilon\left(F_\varepsilon^{-1}(\tau_k)\right)} \right) \end{bmatrix} + o_p(1),$$

and the two-step estimator $\widehat{\theta}(\tau)$ based on (25) has asymptotic representation:

$$\sqrt{n}\left(\widehat{\theta}(\tau) - \theta(\tau)\right) = -\frac{1}{f_\varepsilon\left(F_\varepsilon^{-1}(\tau)\right)} \Omega^{-1} \left\{ \frac{1}{\sqrt{n}} \sum_t Z_t \psi_\tau(u_{t\tau}) \right\}$$
$$+ \Omega^{-1} \Gamma \sqrt{n}\,(\widetilde{a} - a) + o_p(1),$$

where $a = [a_1, a_2, \ldots, a_m]^\top$, $\Omega = E\left[Z_t Z_t{}^\top / \sigma_t\right]$, and

$$\Gamma = \sum_{k=1}^{p} \theta_k C_k,\ C_k = \mathrm{E}\left[\left(|u_{t-k-1}|, \ldots, |u_{t-k-m}| \right) \frac{Z_t}{\sigma_t} \right].$$

Remark 3. Note that the infeasible estimator $\widetilde{\theta}(\tau)$ based on unobserved regressors z_t has the following Bahadur representation:

$$\sqrt{n}\left(\widetilde{\theta}(\tau) - \theta(\tau)\right) = -\frac{1}{f_\varepsilon\left(F_\varepsilon^{-1}(\tau)\right)} \Omega^{-1} \left\{ \frac{1}{\sqrt{n}} \sum_t z_t \psi_\tau(u_{t\tau}) \right\} + o_p(1),$$

we see that the Bahadur representation (and thus the variance) of $\widehat{\theta}(\tau)$ contains an additional term that arises from the preliminary estimation. □

Remark 4. The estimation procedure also provides a robust estimator for the conditional volatility. □

Remark 5. Quantile regression estimation can also be applied to other types of ARCH and GARCH models, say the quadratic ARCH or GARCH models, or the Threshold ARCH/GARCH models based on nonlinear quantile regressions. □

4. Quantile regressions with dependent errors

Quantile regression can also be applied to regression models with dependent errors. Consider the following linear model:

$$Y_t = \alpha + \beta' X_t + u_t = \theta' Z_t + u_t, \tag{27}$$

where X_t and u_t are k and 1-dimensional weakly dependent stationary random variables, $\{X_t\}$ and $\{u_t\}$ are independent with each other, $\mathrm{E}(u_t) = 0$. If we denote distribution function of u_t as $F_u(\cdot)$, then conditional on X_t, the τth quantile of Y_t is given by

$$Q_{Y_t}(\tau|X_t) = \alpha + \beta' X_t + F_u^{-1}(\tau) = \theta(\tau)' Z_t,$$

where $\theta(\tau) = (\alpha + F_u^{-1}(\tau), \beta')'$. The vector of parameters, $\theta(\tau)$, can be estimated by solving the problem

$$\widehat{\theta}(\tau) = \arg\min_{\theta \in \mathbb{R}^p} \sum_{t=1}^{n} \rho_\tau(Y_t - Z_t\theta). \tag{28}$$

Let $u_{t\tau} = Y_t - \theta(\tau)' Z_t$, we have $E[\psi_\tau(u_{t\tau})|X_t] = 0$. Under assumptions on moments and weak dependence of (X_t, u_t),

$$n^{-1/2} \sum_{t=1}^{n} Z_t \psi_\tau(u_{t\tau}) = \begin{bmatrix} n^{-1/2} \sum_{t=1}^{n} \psi_\tau(u_{t\tau}) \\ n^{-1/2} \sum_{t=1}^{n} X_t \psi_\tau(u_{t\tau}) \end{bmatrix} \Rightarrow N(0, \Sigma(\tau)),$$

where $\Sigma(\tau)$ is the long-run covariance matrix of $Z_t \psi_\tau(u_{t\tau})$ defined by

$$\Sigma(\tau) = \lim \left(n^{-1/2} \sum_{t=1}^{n} Z_t \psi_\tau(u_{t\tau}) \right) \left(n^{-1/2} \sum_{t=1}^{n} Z_t \psi_\tau(u_{t\tau}) \right) = \begin{bmatrix} \omega_\psi^2(\tau) & 0 \\ 0 & \Omega(\tau) \end{bmatrix}.$$

Under regularity assumptions, the quantile regression estimator (28) has the following asymptotic representation:

$$\sqrt{n}(\widehat{\theta}(\tau) - \theta(\tau)) = \frac{1}{2f(F^{-1}(\tau))} \Sigma_z^{-1} \frac{1}{n^{1/2}} \sum_{t=1}^{n} Z_t \psi_\tau(u_{t\tau}),$$

where

$$\Sigma_z = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} Z_t Z_t^\top.$$

As a result,

$$\sqrt{n}(\widehat{\theta}(\tau) - \theta(\tau)) \Rightarrow N\left(0, \frac{1}{4f(F^{-1}(\tau))^2} \Sigma_z^{-1} \Sigma(\tau) \Sigma_z^{-1}\right).$$

The above results may be extended to the case where other elements in $\theta(\tau)$ are also τ-dependent.

Statistical inference based on $\widehat{\theta}(\tau)$ requires estimation of the covariance matrices Σ_z and $\Sigma(\tau)$. The matrix Σ_z can be easily estimated by its sample analogue

$$\widehat{\Sigma}_z = n^{-1} \sum_{t=1}^{n} Z_t Z_t^\top,$$

$\Sigma(\tau)$ may be estimated following the HAC estimation literature (see, e.g., Andrews (1991)). Let $\widehat{u}_{t\tau} = Y_t - \widehat{\theta}(\tau)' Z_t$, we may estimate $\Sigma(\tau)$ by

$$\widehat{\Sigma}(\tau) = \sum_{h=-M}^{M} k\left(\frac{h}{M}\right)\left[\frac{1}{n}\sum_{1\le t,t+h\le n} Z_t \psi_\tau(\widehat{u}_{t\tau}) Z_{t+h}^{\top} \psi_\tau(\widehat{u}_{(t+h)\tau})\right],$$

where $k(\cdot)$ is the lag window defined on $[-1, 1]$ with $k(0) = 1$ and M is the bandwidth parameter satisfying the property that $M \to \infty$ and $M/n \to 0$ as the sample size $n \to \infty$.

Portnoy (1991) studied the asymptotic properties for regression quantiles with m-dependent errors; his analysis also allows for nonstationarity with a nonvanishing bias term.

The above quantile regression analysis can also be extended to the case with long-range dependent errors. Koul and Mukherjee (1994) considered linear model (27) when the errors are a function of Gaussian random variables that are stationary and long-range dependent, so that

$$Cov(u_t, u_{t+h}) = h^{-\lambda} L(h), \quad \text{for some } 0 < \lambda < 1,$$

where $L(h)$ is positive for large h and slow varying at infinity.

5. Nonparametric and semiparametric QR models

One direction that has attracted a lot of research attention is the nonparametric and semiparametric time series quantile regression models – see, e.g., Koenker et al. (1994), Honda (2000), Cai (2002), Cai and Xu (2009), Cai and Xiao (2010), and Wei et al. (2006).

5.1. Nonparametric dynamic quantile regressions

Consider the model

$$Q_{Y_t}(\tau|X_t) = \theta_\tau(X_t),$$

where $\theta_\tau(\cdot)$ is an unknown function. We may estimate the conditional quantile function $\theta_\tau(x) = Q_{Y_t}(\tau|X_t = x)$ via nonparametric smoothing. In particular, given time series observations $\{(Y_t, X_t)\}_{t=1}^n$, we may consider the following Nadaraya-Watson nonparametric quantile regression that minimizes the following objective function

$$\widehat{\theta}_\tau(x) = \arg\min_\theta \sum_{t=1}^{n} K_h(X_t - x)\,\rho_\tau(Y_t - \theta),$$

where $K_h(X_t - x) = K((X_t - x)/h)$ and $K(\cdot)$ is a product kernel of $k(\cdot)$, which is symmetric and has a compact support, say $[-1, 1]$, $h = h(n) \to 0$ is a bandwidth

parameter that controls how "close" X_t is from x. Denote $\theta_\tau(x) = \theta_0$, let $f_X(x)$ be the density of X, and $f_{Y|X}(y)$ be the conditional density of Y given X,

$$\mu_j = \int u^j K(u)du, \quad \text{and} \quad \nu_0 = \int K^2(u)du, \tag{29}$$

and let $v = \sqrt{nh^q}(\theta - \theta_\mathbf{0})$, where q is the dimension of X; under appropriate assumptions, we may approximate

$$\sum_{t=1}^{n} K_h(X_t - x)\rho_\tau(Y_t - \theta) - \sum_{t=1}^{n} K_h(X_t - x)\rho_\tau(Y_t - \theta_0)$$

by a quadratic function

$$-\frac{1}{\sqrt{nh^q}} v \left[\sum_{t=1}^{n} K_h(X_t - x)\psi_\tau(u_{t\tau})\right] + \frac{1}{2} f_X(x) f_{Y|X}(Q_Y(\tau|x)) v^2$$

whose minimizer is asymptotically normal and then show that the QR estimator is close enough to the minimizer. The NW estimator of the conditional quantile function $Q_{Y_t}(\tau|X_t = x)$ has the following local Bahadur representation:

$$\sqrt{nh^q}\left(\widehat{\theta}_\tau(x) - \theta_\tau(x)\right) = \frac{1}{f_X(x) f_{Y|X}(Q_Y(\tau|x))} \times \left[\frac{1}{\sqrt{nh^q}} \sum_{t=1}^{n} K_h(X_t - x)\psi_\tau(Y_t - \theta_\tau(x))\right] + o_p(1).$$

If we choose bandwidth h so that $nh^q \to \infty$, $h \to 0$,

$$\sqrt{nh^q}\left(\widehat{\theta}_\tau(x) - \theta_\tau(x) - h^2 B_\tau(x)\right) \Rightarrow N\left(0, \frac{\tau(1-\tau)\nu_0}{f_X(x) f_{Y|X}(Q_Y(\tau|x))^2}\right),$$

where $B_\tau(x)$ is the bias term.

Other types of nonparametric estimators, such as the local polynomial estimator, can also be analyzed in a similar way. Under smoothness condition of $\theta_\tau(\cdot)$, so that it has $(m+1)$th continuous derivative $(m \geq 1)$, for any given point x, when X_t is in a neighborhood of x, $\theta_\tau(X_t)$ can be approximated by a polynomial function as

$$\theta_\tau(X_t) \approx \theta_\tau(x) + \theta'_\tau(x)(X_t - x) + \cdots + \theta^{(m)}_\tau(x)(X_t - x)^m/m!,$$

thus

$$Q_{Y_t}(\tau|X_t) \approx \sum_{j=0}^{m} \theta_{j\tau}^T (X_t - x)^j,$$

where $\theta_{j\tau} = \theta_\tau^{(j)}(x)/j$ for $0 \le j \le m$. Then, we may estimate $\theta_\tau(x)$ based on

$$\min_\theta \sum_{t=1}^{n} K_h\left(X_t - x\right) \rho_\tau \left(Y_t - \sum_{j=0}^{m} \theta_j^T \ \left(X_t - x\right)^j \right).$$

Like the nonparametric mean regressions, the nonparametric quantile regression estimator suffers the "curse of dimensionality." Various dimension reduction methods have been proposed in the literature, including additive nonparametric models and functional coefficient quantile regressions. Cai and Xu (2009) studied the dynamic functional coefficient quantile regression models, extending the results of Honda (2004) to the time series case.

Consider a stationary sequence $\{Y_t, X_t, Z_t\}_{t=-\infty}^{\infty}$, let

$$Q_{Y_t}(\tau|x,z) = Q_{Y_t}(\tau|(X_t, Z_t) = (x,z))$$

be the conditional quantile function of Y_t given $(X_t, Z_t) = (x,z)$, for any $0 < \tau < 1$, a functional (or, varying) coefficient quantile regression model takes the following form:

$$Q_{Y_t}(\tau|X_t, Z_t) = \alpha_\tau(X_t)^\top Z_t. \tag{30}$$

Under smoothness condition of coefficient functions $\alpha_\tau(\cdot)$, if X_t is in a neighborhood of x, we have

$$Q_{Y_t}(\tau|X_t, Z_t) = \alpha_\tau(X_t)^\top Z_t \approx \sum_{j=0}^{m} \theta_{j\tau}^T \ Z_t \left(X_t - x\right)^j,$$

where $\theta_{j\tau} = \alpha_\tau^{(j)}(x)/j!$ for $0 \le j \le m$. Then, we may estimate $\alpha_\tau(x)$ based on the following local polynomial functional coefficient quantile regression estimation

$$\min_\theta \sum_{t=1}^{n} K_h\left(X_t - x\right) \rho_\tau \left(Y_t - \sum_{j=0}^{m} \theta_j^T \ Z_t \left(X_t - x\right)^j \right).$$

Under regularity assumptions, Cai and Xu (2009) show that

$$\sqrt{nh^q}\left(\widehat{\alpha}_\tau(x) - \alpha_\tau(x) - h^2 b_\tau(x)\right) \Rightarrow N\left(0, \frac{\tau(1-\tau)\nu_2(K)}{f_X(x)} \ \Omega^*(x)^{-1}\Omega(x)\Omega^*(x)^{-1}\right).$$

where $b_\tau(x) = \frac{1}{2}\left(\int u^2 K(u)du\right)\alpha_\tau''(x) + o_p(1)$, and

$$\Omega^*(x) = \mathrm{E}\left[f_{Y|X,Z}\left(Q_Y\left(\tau|X_t, Z_t\right)\right) Z_t Z_t^\top | X_t = x\right], \Omega(x) = E\left[Z_t Z_t^\top | X_t = x\right].$$

Koenker et al. (1994) proposed nonparametric quantile regression method via smoothing splines. They consider a class of quantile smoothing splines defined as a solution of

$$\min_{g \in \mathcal{G}} \sum_{t=1}^{n} \rho_\tau\left(Y_t - g\left(X_t\right)\right) + \lambda V(g'), \tag{31}$$

for appropriately chosen $\mathcal{G}$, where $V(f)$ denote the total variation norm of f, and λ is the smoothing parameter. If g' is sufficiently smooth, $V(g') = \int \left|g''(x)\right| dx$. Koenker et al. (1994) show that the solutions $\widehat{g}(x)$ of (31) are continuous, piecewise linear functions. For λ sufficiently large, the solution is the corresponding globally linear regression quantile. Computation of quantile smoothing splines can be efficiently carried out by linear programming methods.

Chen and Shen (1998) and Chen (2006) studied general sieve estimates for weakly dependent data that can be applied to the estimation of quantiles.

Instead of using the "check" function, an alternative nonparametric approach to estimate conditional quantiles is to invert an estimator of the conditional distribution function. Yu and Jones (1998) propose first estimating the conditional distribution function using the "double-kernel" local linear technique of Fan et al. (1996) and then inverting the conditional distribution estimator to obtain an estimator of a conditional quantile. The Yu and Jones (1998) estimator has nice properties such as no boundary effects and design adaptation, but it produces conditional distribution estimators that are not constrained to lie within [0,1] or be monotonic, and modifications are needed. Cai (2002) proposed another estimator for the conditional quantile by inverting the weighted Nadaraya-Watson estimator of the conditional distribution function of Hall et al. (1999). Giving a stationary strong mixing process $\{Y_t, X_t\}$, the weighted Nadaraya-Watson estimator of the conditional distribution function is defined as

$$\widehat{F}(y|x) = \frac{\sum_{t=1}^{n} p_t(x) K\left(\frac{x-X_t}{h}\right) I(Y_t \leq y)}{\sum_{t=1}^{n} p_t(x) K\left(\frac{x-X_t}{h}\right)},$$

where $p_t(x) \geq 0$ are the weighting functions and $\sum_{t=1}^{n} p_t(x) = 1$. The weighted Nadaraya-Watson estimator has nice properties such as being in [0,1] and monotonic increasing, and good boundary behavior. Cai (2002) proposes the following estimate for the τth conditional quantile of Y_t:

$$\widehat{Q}_{Y_t}(\tau|x) = \inf\left\{y : \widehat{F}(y|x) \geq \tau\right\}.$$

Under smoothness conditions on the densities and mixing conditions that controls the dependence, Cai (2002) shows that $\widehat{Q}_{Y_t}(\tau|x)$ is a consistent estimator for $Q_{Y_t}(\tau|X_t = x)$, and, letting $f(y|x)$ be the conditional density function and $f_X(\cdot)$ be the marginal density of X,

$$\sqrt{nh}\left[\widehat{Q}_{Y_t}(\tau|x) - Q_{Y_t}(\tau|x) - h^2 B_\tau(x) + o_p(h^2)\right] \xrightarrow{d} \mathcal{N}\left(0, \sigma_\tau^2(x)\right),$$

where

$$B_\tau(x) = -\frac{1}{2}\mu_2 \frac{\partial^2 F\left(Q_Y(\tau|x)|x\right)/\partial x^2}{f(Q_Y(\tau|x)|x)}, \quad \sigma_\tau^2(x) = \frac{\tau(1-\tau)\nu_0}{f(Q_Y(\tau|x)|x)^2 f_X(x)}.$$

5.2. *Semiparametric dynamic quantile regressions*

Following the partially linear approach in conditional mean models, Cai and Xiao (2010) consider another dimension-reduction modeling method – the partially varying

coefficient models. The partially varying coefficient quantile regression model serves as an intermediate class of models with good robustness by nonparametric treatment on certain covariates and relatively more precise estimation on the parametric effect of other variables. In this semiparametric approach, existing information concerning possible linearity of some of the components can be taken into account in such models to improve efficiency.

A partially varying coefficient quantile regression model for time series data takes the following semiparametric form,

$$Q_{Y_t}(\tau|X_t, Z_t) = \beta_\tau^T Z_{t1} + \alpha_\tau(X_t)^T Z_{t2},$$

where $Z_t = (Z_{t1}^T, Z_{t2}^T)^T \in \Re^{p+q}$, $\alpha_\tau(\cdot) = (a_{1,\tau}(\cdot), \ldots, a_{q,\tau}(\cdot))^T$, and $\{a_{k,\tau}(\cdot)\}$ are smooth coefficient functions. Given this model, if β_τ were known, we would be able to construct the following partial quantile residual: $Y_{t1} = Y_t - \beta_\tau^T Z_{t1}$, thus

$$Q_{Y_{1t}}(\tau|X_t, Z_{2t}) = \alpha_\tau(X_t)^T Z_{t2}.$$

Then, one may estimate $\alpha_\tau(u_0)$ based on the nonparametric functional coefficient quantile regression estimation.

In practice, β_τ is unknown. To estimate both the parameter vector β and the functional coefficients $\alpha(\cdot)$, we may first treat β as a function of X_t, $\beta(X_t)$, then the model becomes a functional coefficient model, and all coefficient functions can be estimated by using the following local fitting,

$$\min_{\beta,\theta} \sum_{t=1}^{n} K_h(X_t - x)\, \rho_\tau \left(Y_t - \beta^T Z_{t1} - \sum_{j=0}^{m} \theta_j^T Z_{t2} (X_t - x)^j \right). \tag{32}$$

We denote the above local polynomial estimator of β as $\widehat{\beta}(x)$. Notice that although β is a global parameter, the above estimation of β involves only local data points in a neighborhood of x, so that it is not optimal. Indeed, $\widehat{\beta}(\cdot) - \beta = O_p((n h^q)^{-1/2})$. An optimal estimation of the constant coefficients requires using all data points, and the optimal convergence rate should be $\sqrt{n}$ instead of $\sqrt{n h^q}$. To obtain a $\sqrt{n}$-consistent estimator for β_τ, we may use the following averaging method to obtain a second-stage estimator of β that achieves the optimal rate of convergence:

$$\widetilde{\beta} = \widetilde{\beta}_\tau = \frac{1}{n} \sum_{t=1}^{n} \widehat{\beta}(X_t). \tag{33}$$

To estimate the functional coefficients $\alpha(\cdot)$, we define the estimated partial quantile residual as $Y_{t*} = Y_t - \widetilde{\beta}^T Z_{t1}$, where $\widetilde{\beta}$ is a $\sqrt{n}$-consistent estimate of β, and consider the following feasible local polynomial functional coefficient estimation:

$$\min_{\theta} \sum_{t=1}^{n} K_{h_1}(X_t - x)\, \rho_\tau \left(Y_{t*} - \sum_{j=0}^{m} \theta_j{}^T Z_{t2} (X_t - x)^j \right), \tag{34}$$

where h_1 is the bandwidth used for this step, which is different from the bandwidth used in (32). Solving the minimization problem in (34) gives $\widetilde{\alpha}(u_0) = \widehat{\theta}_{0*}$, the local polynomial estimate of $\alpha(u_0)$, and $\widetilde{\alpha}^{(j)}(u_0) = j\,!\,\widehat{\theta}_{j*}$ $(j \geq 1)$, the local polynomial estimate of the jth derivative $\alpha^{(j)}(u_0)$ of $\alpha(u_0)$.

Choosing $h/h_1 = o(1)$, and under other regularity assumptions, the above nonparametric estimator is "oracle," in the sense that the asymptotic properties of this nonparametric estimator are not affected by preliminary estimation of β_τ. Denote $f_X(\cdot)$, the marginal density of X_t, and $f_{y|z,x}(\cdot|\cdot)$, the conditional density of Y_t, given (Z_t, X_t), let

$$\Omega(x) = E\left[Z_t Z_t^T | X_t = x\right] \quad \text{and} \quad \Omega^*(x) = E\left[Z_t Z_t^T f_{y|z,x}(q_\tau(Z_t, X_t))|X_t = x\right],$$

$$B_1^* = e_1^T\, E\left[(\Omega^*(X_1))^{-1}\Omega^{*'}(X_1)\begin{pmatrix}0\\ \alpha'(X_1)\end{pmatrix}\right],$$

where $\Omega^{*'}(x)$ is the first-order derivative of $\Omega^*(x)$ and $e_1^T = (I_p, 0_{p\times q})$ with I_p being a $p \times p$ identity matrix and $0_{p\times q}$ being a $p \times q$ zero matrix, and $B_2^* = e_1^T\, E\left[(\Omega^*(X_1))^{-1}\Gamma(X_1)\right]$, where

$$\Gamma(x) = E\left[f'_{y|z,x}(q_\tau(Z_t, X_t))Z_t\left(\alpha'(X_t)^T Z_{t2}\right)^2 |X_t = x\right],$$

and $f'_{y|z,x}(y)$ denotes the derivative of $f_{y|z,x}(y)$ with respect to y. Under regularity assumptions,

$$\sqrt{n}\left[\widetilde{\beta}_\tau - \beta_\tau - B_\beta\right] \xrightarrow{d} \mathcal{N}(0,\ \Sigma_\beta),$$

where the asymptotic bias term is $B_\beta = h^2\mu_2(B_1^* - B_2^*/2)$, μ_2 is defined as (29), and the asymptotic variance is

$$\Sigma_\beta = \tau(1-\tau)\, E\left[e_1^T(\Omega^*(X_1))^{-1}\Omega(X_1)(\Omega^*(X_1))^{-1}e_1\right]$$
$$+ 2\sum_{s=1}^{\infty}\mathrm{Cov}(e_1^T\left(\Omega^*(X_1)\right)^{-1}Z_1\,\eta_1, e_1^T(\Omega^*(X_{s+1}))^{-1}Z_{s+1}\,\eta_{s+1}).$$

Here, $\eta_t = \tau - I\{Y_t \leq Q_Y(\tau|Z_t, X_t)\}$.

Linton and Shang (2010) studied the conditional quantile estimation in a semiparametric GARCH model. In particular, they consider the following quadratic GARCH(1,1) model:

$$u_t = \sigma_t \cdot \varepsilon_t, \sigma_t^2 = \gamma_0 + \beta_1\sigma_{t-1}^2 + \gamma_1 u_{t-1}^2,$$

where ε_t are i.i.d. (0,1). The conditional quantile is given by $Q_{u_t}(\tau|\mathcal{F}_{t-1}) = \sigma_t F_\varepsilon^{-1}(\tau)$. Linton and Shang (2010) studied efficient estimation of the GARCH parameters $(\gamma_0, \beta_1, \gamma_1)$ and nonparametric estimation of $F_\varepsilon^{-1}(\tau)$ based on inverting the distribution estimator. Notice that ε_t is standardized to be variance 1, they consider a weighted

empirical distribution estimator for $F_\varepsilon(\cdot)$, where the weights are determined by

$$\{\widehat{w}_t\} = \arg\max_{w_t} \left\{\Pi_{t=1}^n w_t\right\}$$
$$\text{s.t.} \sum_{t=1}^n w_t = 1; \sum_{t=1}^n w_t \varepsilon_t = 0; \sum_{t=1}^n w_t \left(\varepsilon_t^2 - 1\right) = 0.$$

The weighted empirical distribution estimator for $F_\varepsilon(\cdot)$ is then given by

$$\widehat{F}_\varepsilon(\cdot) = \sum_{t=1}^n \widehat{w}_t 1\left(\varepsilon_t \le x\right),$$

and $F_\varepsilon^{-1}(\tau)$ is estimated by $\widehat{F}_\varepsilon^{-1}(\tau) = \sup\left\{s : \widehat{F}_\varepsilon(s) \le \tau\right\}$. Also see Komunjer and Vuong (2010) for semiparametric estimations based on minimizing an M-objective function.

6. Other dynamic quantile models

6.1. The CAViaR model and local modeling methods

Quantile-based method provides a local approach to directly model the dynamics of a time series at a specified quantile.

Consider again the linear GARCH model given by (17) and (18). Note that $\sigma_{t-j} F^{-1}(\tau) = Q_{u_{t-j}}(\tau|\mathcal{F}_{t-j-1})$; hence, the conditional quantile $Q_{u_t}(\tau|\mathcal{F}_{t-1})$ has the following representation:

$$Q_{u_t}(\tau|\mathcal{F}_{t-1}) = \beta_0^* + \sum_{i=1}^p \beta_i^* Q_{u_{t-i}}(\tau|\mathcal{F}_{t-i-1}) + \sum_{j=1}^q \gamma_j^* \left|u_{t-j}\right|, \tag{35}$$

where $\beta_0^* = \beta_0(\tau) = \beta_0 F^{-1}(\tau)$, $\beta_i^* = \beta_i$, $i = 1, \ldots, p$, $\gamma_j^* = \gamma_j(\tau) = \gamma_j F^{-1}(\tau)$, and $j = 1, \ldots, q$. From (35) we can see an important feature of the linear GARCH model: conditional quantiles $Q_{u_t}(\tau|\mathcal{F}_{t-1})$ themselves follow an autoregression. This representation suggests that one may model the local dynamics or local correlation directly based on the conditional quantiles.

Engle and Mangenelli (2004) propose the Conditional Autoregressive Value-at-Risk (CAViaR) specification for the τth conditional quantile of u_t:

$$Q_{u_t}(\tau|\mathcal{F}_{t-1}) = \beta_0 + \sum_{i=1}^p \beta_i Q_{u_{t-i}}(\tau|\mathcal{F}_{t-i-1}) + \sum_{j=1}^q \alpha_j \ell(X_{t-j}), \tag{36}$$

where $X_{t-j} \in \mathcal{F}_{t-j}$, $\mathcal{F}_{t-j}$ is the information set at time $t-j$. A natural choice of X_{t-j} is the lagged u. When we choose $X_{t-j} = \left|u_{t-j}\right|$, we obtain (35). Engle and Mangenelli (2004) discussed many choices of $\ell(X_{t-j})$ leading to different specifications of the CAViaR model.

Sim and Xiao (2009) and Sim (2009) considered local models to study the asymmetric correlation of international stock returns. To study the correlations between the τ_Yth quantile of Y_t and the τ_Xth quantile of X_t, they consider the following quantile dependence model:

$$Q_{Y_t}(\tau_Y|\mathcal{F}_{t-1}) = h(Q_{X_t}(\tau_X|V_t), \beta(\tau_X, \tau_Y)). \tag{37}$$

Let $X_t = Y_{t-1}$, $\tau_X = \tau_Y = \tau$, and $Q_{X_t}(\tau_X|V_t) = Q_{X_t}(\tau_X|\mathcal{F}_{t-2})$, and consider linear function of $h(\cdot)$, we obtain an autoregression model for the τth conditional quantile of Y_t:

$$Q_{y_t}(\tau|\mathcal{F}_{t-1}) = \beta_0 + \beta Q_{y_{t-1}}(\tau|\mathcal{F}_{t-2}).$$

The model can be extended to include additional regressors. For instance, let $\tau_X = \tau_Y = \tau$, and consider the following quantile model

$$Q_{Y_t}(\tau|\mathcal{F}_{t-1}) = h(Q_{Z_{t-1}}(\tau|\mathcal{F}_{t-2}), Q_{Y_{t-1}}(\tau|\mathcal{F}_{t-2}), \beta(\tau)),$$

where the τth conditional quantile of Y_t is affected by its own lagged value and lagged values of the conditional quantile of covariates.

Estimation of the CAViaR model is challenging. If we denote the vector of unknown parameters by θ, and, for simplicity, denote $Q_{u_t}(\tau|\mathcal{F}_{t-1})$ by $Q_t(\tau, \theta)$, then we may consider estimate θ by minimizing:

$$RQ_n(\tau, \theta) = \sum_t \rho_\tau(Y_t - Q_t(\tau, \theta)), \tag{38}$$

where $Q_t(\tau, \theta) = \beta_0 + \sum_{i=1}^p \beta_i Q_{t-i}(\tau, \theta) + \sum_{j=1}^q \alpha_j \ell(X_{t-j})$. Because conditional quantiles enters the CAViaR regression model as regressors and they are latent, conventional nonlinear quantile regression techniques are not directly applicable. De Rossi and Harvey (2009) studied an iterative Kalman filter method to calculate dynamic conditional quantiles that may be applied to calculate certain types CAViaR models. In their model, the observed time series Y_t is described by measurement equation

$$Y_t = \xi_t(\tau) + \varepsilon_t(\tau),$$

where $\xi_t(\tau) = Q_{Y_t}(\tau|\mathcal{F}_{t-1})$ is the state variable and the disturbances $\varepsilon_t(\tau)$ are assumed to be serially independent and independent of $\xi_t(\tau)$. The dynamics of this system is characterized by the state transition equation based on $\xi_t(\tau)$. For example, if the conditional quantiles follow an autoregression, we have

$$\xi_t(\tau) = \beta\xi_{t-1}(\tau) + \eta_t(\tau).$$

Alternative forms of the state transition equations can be considered. The above state-space model can then be estimated by iteratively applying an appropriate signal extraction algorithm.

Hsu (2010) studied estimating CAViaR with an MCMC method based on the Bayes approach of Yu and Moyeed (2001). She considered the following asymmetric Laplace density as a working conditional density for the error term in the CAViaR model:

$$f(\varepsilon_{t\tau}|\mathcal{F}_{t-1}) = \frac{\tau(1-\tau)}{\sigma}\exp\left\{-\frac{1}{\sigma}\rho_\tau(\varepsilon_{t\tau})\right\},$$

where $\varepsilon_{t\tau} = Y_t - Q_t(\tau,\theta)$ and σ is a scale parameter, then, given a data set of size n, the working likelihood is

$$f(\text{Data}|\theta,\sigma) = \left(\frac{\tau(1-\tau)}{\sigma}\right)^n \exp\left\{-\frac{1}{\sigma}RQ_n(\tau,\theta)\right\},$$

where $RQ_n(\tau,\theta)$ is given by (38). Hsu (2010) choose a flat prior for each coefficient in θ and the inverse gamma distribution $\text{IG}(\alpha_0, s_0)$ for σ, thus the joint prior for θ is given by

$$\pi(\theta,\sigma) \propto \frac{1}{\sigma^{\alpha_0+1}}\exp\left(-\frac{s_0}{\sigma}\right),$$

and the posterior for (θ,σ) is

$$f(\theta,\sigma|\text{Data}) \propto \frac{1}{\sigma^{\alpha_0+1}}\left(\frac{\tau(1-\tau)}{\sigma}\right)^n \exp\left\{-\frac{1}{\sigma}RQ_n(\tau,\theta) - \frac{s_0}{\sigma}\right\}.$$

Posterior inference on the CAViaR model can then be implemented.

6.2. *Additive quantile models*

Gourieroux and Jasiak (2008) proposed a dynamic additive quantile model based on a group of baseline quantile functions. In particular, let $Q_{Y_t}(\tau|\mathcal{F}_{t-1})$ be dependent on some unknown parameters θ and denote it as $Q_{Y_t|X_t}(\tau;\theta)$, they define a dynamic additive quantile model as

$$Q_{Y_t|X_t}(\tau;\theta) = \sum_{k=1}^{K}\rho_k(X_t,\alpha_k)\,Q_k(\tau,\beta_k) + \rho_0(X_t,\alpha_0),$$

where $Q_k(\tau,\beta_k)$ are baseline quantile functions with identical range and $\rho_k(X_t,\alpha_k)$ are positive functions of the past information. By construction, the quantile curves do not cross. Information-based estimation methods are proposed by Gourieroux and Jasiak (2008) to estimate these models.

6.3. *QR for dynamic panel*

Galvao (2010) recently studied quantile regression with dynamic panel data. In particular, he considered the following dynamic panel quantile model

$$\begin{aligned} Q_{Y_{it}}(\tau|Z_{it},Y_{i,t-1},X_{it}) &= Z_{it}\eta(\tau) + \alpha(\tau)\,Y_{i,t-1} + X_{it}'\beta(\tau), \\ &\qquad i=1,\ldots,n; \quad t=1,\ldots,T, \end{aligned} \tag{39}$$

where Z_{it} takes value 0 or 1 that identifies the fixed effects for the n groups and $\eta = (\eta_1, \ldots, \eta_n)^\top$ is the $n \times 1$ vector of individual specific effects. In the presence of lagged y, a direct quantile regression based on (39) that minimizes

$$\sum_{i=1}^{n}\sum_{t=1}^{T} \rho_\tau \left(Y_{it} - Z_{it}\eta - \alpha Y_{i,t-1} - X_{it}'\beta\right)$$

is potentially biased. Galvao studied the instrumental variable estimation for the above dynamic panel mode. Assuming that there is instrumental variable W_{it} that affects $Y_{i,t-1}$ but are independent of the errors, following Chernozhukov and Hansen (2008), he considers: for fixed α,

$$(\widehat{\eta}(\alpha), \widehat{\beta}(\alpha), \widehat{\gamma}(\alpha)) = \arg\min_{\eta,\beta,\gamma} \sum_{i=1}^{n}\sum_{t=1}^{T} \rho_\tau \left(Y_{it} - Z_{it}\eta - \alpha Y_{i,t-1} - X_{it}'\beta - W_{it}'\gamma\right)$$

and estimates α by solving for

$$\widehat{\alpha} = \min_{\alpha} \|\widehat{\gamma}(\alpha)\|_A,$$

where $\|x\|_A = x'Ax$. The final estimators for $\alpha(\tau)$ and $\beta(\tau)$ are then given by $(\widehat{\alpha}(\tau), \widehat{\beta}(\widehat{\alpha}(\tau), \tau))$.

7. Extremal quantile regressions

Many statistical applications focus on either the lower quantile or upper quantiles of the distribution or conditional distribution. Consequently, theory of extremal quantiles may be used in such applications. Without loss of generality, we consider the lower extreme quantiles (i.e., $\tau \searrow 0$) only.

Given a random sample of n observations $\{Y_t, X_t\}_{t=1}^n$, we are interested in the τth quantile of Y or the τth conditional quantile of Y given X. Knight (2001) and Portnoy and Jureckova (1999) studied asymptotic behavior of extremal quantile regression estimators when $\tau n \to 0$ as $n \to \infty$. In particular, Knight (2001) investigated extremal quantile regression estimators via the point process approach, and Portnoy and Jureckova (1999) studied it using a density convergence approach. Chernozhukov (2005) studied asymptotic behavior of extremal quantile regression estimators when $\tau n \to \kappa$ and $\tau n \to \infty$ as $n \to \infty$. If $\tau \searrow 0$ and $\tau n \to \kappa \geq 1$ as $n \to \infty$, he calls the corresponding quantile an extremal quantile; if $\tau \searrow 0$ and $\tau n \to \infty$ as $n \to \infty$, the corresponding quantile is called an intermediate-order quantile. In these cases, $\tau = \tau(n)$ (and converges to 0 as $n \to \infty$) is a sequence of quantiles index associated with the sample size n.

The limiting behavior of extremal quantiles depends not only on the types of quantiles but also on the tail behavior of the distributions (or conditional distributions). Consider, say, the classical linear quantile regression model, where

$$Q_Y(\tau|X) = X'\theta(\tau), \tag{40}$$

where X is a d-dimensional vector, suppose that there exists an auxiliary parameter θ_0, such that $U = Y - X'\theta_0$ has conditional lower endpoint 0 (or $-\infty$) a.s. and its conditional quantile function $Q_U(\tau|X)$ satisfies the following tail conditions: As $\tau \searrow 0$, uniformly in the support of X,

$$Q_U(\tau|X) = Q_Y(\tau|X) - X'\theta_0 \sim F_U^{-1}(\tau),$$

where $F_U^{-1}(\tau)$ is a quantile function exhibiting Pareto-type behavior in the tails, such that $F_U^{-1}(\tau) \sim L(\tau)\tau^{-\xi}$, where $L(\tau)$ is a slow-varying function at 0. The number ξ is called the extreme value index.

Given time series observations $\{Y_t, X_t\}_{t=1}^n$ and the quantile regression model (40), if we estimate $Q_Y(\tau|X)$ by $X'\widehat{\theta}(\tau)$, where $\widehat{\theta}(\tau)$ is estimated via quantile regression (2), under appropriate assumptions, the extreme quantiles converge to non-normal distributions, and the intermediate-order quantiles converge to normal limits. If we consider the canonically normalized regression quantile

$$Z_n^*(\tau) = \frac{1}{F_U^{-1}(1/n)}\left(\widehat{\theta}(\tau) - \theta(\tau)\right),$$

under the assumption that $\{Y_t, X_t\}_t$ is weakly dependent stationary sequence with extreme events satisfying a nonclustering condition, Chernozhukov (2005) show that if $\tau \searrow 0$ and $\tau n \to \kappa \geq 1$ as $n \to \infty$,

$$Z_n^*(\tau) \Rightarrow Z_\infty(\kappa) - \kappa^{-\xi},$$

where

$$Z_\infty(\kappa) = \arg\min_z \left[-\kappa\mu_X' z + \sum_{i=1}^{\infty}\left[X_i' z - \Gamma_i^{-\xi}\right]_+\right], \xi < 0$$

$$Z_\infty(\kappa) = \arg\min_z \left[-\kappa\mu_X' z + \sum_{i=1}^{\infty}\left[X_i' z + \Gamma_i^{-\xi}\right]_+\right], \xi > 0$$

where $\{\Gamma_1, \Gamma_2, \ldots\} = \{\mathcal{E}_1, \mathcal{E}_1 + \mathcal{E}_2, \ldots\}$ and $\{\mathcal{E}_1, \mathcal{E}_2, \ldots\}$ is an i.i.d. sequence of exponential variables that is independent of $\{X_1, X_2, \ldots\}$, $\mu_X = E(X)$.

The canonically normalized regression quantile is infeasible due to the standardization by $F_U^{-1}(1/n)$. As an alternative, one may consider the self-normalized regression quantile

$$Z_n(\kappa) = \frac{\sqrt{\tau n}}{\overline{X}'\left(\widehat{\theta}(m\tau) - \widehat{\theta}(\tau)\right)}\left(\widehat{\theta}(\tau) - \theta(\tau)\right),$$

for any m, such that $\kappa(m-1) > d$, Chernozhukov (2005) shows that, again, under weak dependence and nonclustering conditions, if $\tau \searrow 0$ and $\tau n \to \kappa \geq 1$ as $n \to \infty$,

$$Z_n(\kappa) \Rightarrow \frac{\sqrt{\kappa} Z_\infty(\kappa)}{\mu_X'\left[Z_\infty(m\kappa) - Z_\infty(\kappa)\right]},$$

and, if $\tau \searrow 0$ and $\tau n \to \infty$ as $n \to \infty$,

$$Z_n(\kappa) \Rightarrow N\left(0, \Sigma_X^{-1} \frac{\xi^2}{\left(m^{-\xi} - 1\right)^2}\right),$$

where $\Sigma_X = E(XX')$. The limiting distributions can be approximated via bootstrap or subsampling methods, and thus, statistical inference can be conducted based on such methods. See, e.g., Chernozhukov (2005) for more details.

8. Quantile regression for nonstationary time series

8.1. Unit root quantile regressions

An important model in economic time series analysis is the autoregressive unit root model, where the differenced time series is stationary (I(0)). Quantile regression can also be applied to unit root time series.

One of the most widely used unit root model is the following Augmented Dickey-Fuller (ADF) regression model

$$Y_t = \alpha_1 Y_{t-1} + \sum_{j=1}^{q} \alpha_{j+1} \Delta Y_{t-j} + u_t, \tag{41}$$

where u_t is i.i.d. $(0, \sigma^2)$. Under assumptions that all the roots of $A(L) = 1 - \sum_{j=1}^{q} \alpha_{j+1} L^j$ lie outside the unit circle, if $\alpha_1 = 1$, Y_t contains a unit root; and if $|\alpha_1| < 1$, Y_t is stationary. If we denote the σ-field generated by $\{u_s, s \le t\}$ by $\mathcal{F}_t$, then conditional on $\mathcal{F}_{t-1}$, the τth conditional quantile of Y_t is given by

$$Q_{Y_t}(\tau|\mathcal{F}_{t-1}) = Q_u(\tau) + \alpha_1 Y_{t-1} + \sum_{j=1}^{q} \alpha_{j+1} \Delta Y_{t-j}.$$

Let $\alpha_0(\tau) = Q_u(\tau)$, $\alpha_j(\tau) = \alpha_j$, $j = 1, \ldots, p$, $p = q + 1$, and define

$$\alpha(\tau) = (\alpha_0(\tau), \alpha_1, \ldots, \alpha_{q+1}), \; X_t = (1, Y_{t-1}, \Delta Y_{t-1}, \ldots, \Delta Y_{t-q})',$$

we have $Q_{Y_t}(\tau|\mathcal{F}_{t-1}) = X_t' \alpha(\tau)$. The unit root quantile autoregressive model can be estimated by:

$$\min_{\alpha} \sum_{t=1}^{n} \rho_\tau(Y_t - X_t^\top \alpha).$$

Denote $w_t = \Delta Y_t$, $u_{t\tau} = Y_t - X_t' \alpha(\tau)$, under the unit root hypothesis and other regularity assumptions,

$$n^{-1/2} \sum_{t=1}^{[nr]} (w_t, \psi_\tau(u_{t\tau}))^\top \Rightarrow (B_w(r), B_\psi^\tau(r))^\top = BM(0, \underline{\Sigma}(\tau)),$$

where

$$\underline{\Sigma}(\tau) = \begin{bmatrix} \sigma_w^2 & \sigma_{w\psi}(\tau) \\ \sigma_{w\psi}(\tau) & \sigma_\psi^2(\tau) \end{bmatrix}$$

is the long-run covariance matrix of the bivariate Brownian motion and can be written as $\Sigma_0(\tau) + \Sigma_1(\tau) + \Sigma_1^\top(\tau)$, where $\Sigma_0(\tau) = E[(w_t, \psi_\tau(u_{t\tau}))^\top (w_t, \psi_\tau(u_{t\tau}))]$ and

$$\Sigma_1(\tau) = \sum_{s=2}^{\infty} E[(w_1, \psi_\tau(u_{1\tau}))^\top (w_s, \psi_\tau(u_{s\tau}))].$$

In addition, $n^{-1}\sum_{t=1}^n Y_{t-1}\psi_\tau(u_{t\tau}) \Rightarrow \int_0^1 B_w dB_\psi^\tau$.

The random function $n^{-1/2}\sum_{t=1}^{[nr]} \psi_\tau(u_{t\tau})$ converges to a two-parameter process $B_\psi^\tau(r) = B_\psi(\tau, r)$, which is partially a Brownian motion and partially a Brownian bridge in the sense that for fixed r, $B_\psi^\tau(r) = B_\psi(\tau, r)$ is a rescaled Brownian bridge, while for each τ, $n^{-1/2}\sum_{t=1}^{[nr]} \psi_\tau(u_{t\tau})$ converges weakly to a Brownian motion with variance $\tau(1-\tau)$. Thus, for each fixed pair (τ, r), $B_\psi^\tau(r) = B_\psi(\tau, r) \sim N(0, \tau(1-\tau)r)$. Let $\widehat{\alpha}(\tau) = (\widehat{\alpha}_0(\tau), \widehat{\alpha}_1, \ldots, \widehat{\alpha}_p)$ and $D_n = \mathrm{diag}(\sqrt{n}, n, \sqrt{n}, \ldots, \sqrt{n})$, the limiting distribution of $\widehat{\alpha}(\tau)$ is summarized in the following Theorem (Koenker and Xiao, 2004).

THEOREM 5. *Let y_t be determined by (41), under the unit root assumption $\alpha_1 = 1$, and other regularity conditions,*

$$D_n(\widehat{\alpha}(\tau) - \alpha(\tau)) \Rightarrow \frac{1}{f(F^{-1}(\tau))} \begin{bmatrix} \int_0^1 \overline{B}_w \overline{B}_w^\top & 0_{2\times q} \\ 0_{q\times 2} & \Omega_\Phi \end{bmatrix}^{-1} \begin{bmatrix} \int_0^1 \overline{B}_w dB_\psi^\tau \\ \Phi \end{bmatrix},$$

where $\overline{B}_w(r) = [1, B_w(r)]^\top$, $\Phi = [\Phi_1, \ldots, \Phi_q]^\top$ is a q-dimensional normal variate with covariance matrix $\tau(1-\tau)\Omega_\Phi$, where

$$\Omega_\Phi = \begin{bmatrix} \nu_0 & \cdots & \nu_{q-1} \\ \vdots & \ddots & \vdots \\ \nu_{q-1} & \cdots & \nu_0 \end{bmatrix}, \quad \nu_j = E[w_t w_{t-j}],$$

and Φ is independent with $\int_0^1 \overline{B}_w dB_\psi^\tau$.

As an immediate by-product of the above Theorem, the limiting distribution of $n(\widehat{\alpha}_1(\tau) - 1)$ is invariant to the estimation of $\widehat{\alpha}_j(\tau)(j = 2, \ldots, p)$ and the lag length p. In particular,

$$n(\widehat{\alpha}_1(\tau) - 1) \Rightarrow \frac{1}{f(F^{-1}(\tau))} \left[\int_0^1 \underline{B}_w^2\right]^{-1} \int_0^1 \underline{B}_w dB_\psi^\tau, \tag{42}$$

where $\underline{B}_w(r) = B_w(r) - \int_0^1 B_w$ is a demeaned Brownian motion.

Inference based on the autoregression quantile process provides a robust approach to testing the unit root hypothesis. Like the conventional ADF t-ratio test, we may consider the t-ratio statistic

$$t_n(\tau) = \frac{f(\widehat{F^{-1}(\tau)})}{\sqrt{\tau(1-\tau)}} \left(Y_{-1}^\top P_X Y_{-1}\right)^{1/2} (\widehat{\alpha}_1(\tau) - 1),$$

where $f(\widehat{F^{-1}(\tau)})$ is a consistent estimator of $f(F^{-1}(\tau))$, Y_{-1} is the vector of lagged dependent variables (Y_{t-1}) and P_X is the projection matrix onto the space orthogonal to $X = (1, \Delta Y_{t-1}, \ldots, \Delta Y_{t-q})$. Under the unit root hypothesis, we have

$$t_n(\tau) \Rightarrow t(\tau) = \frac{1}{\sqrt{\tau(1-\tau)}} \left[\int_0^1 \underline{B}_w^2\right]^{-1/2} \int_0^1 \underline{B}_w dB_\psi^\tau. \tag{43}$$

At any fixed τ, the test statistic $t_n(\tau)$ is simply the quantile regression counterpart of the well-known ADF t-ratio test for a unit root. The limiting distribution of $t_n(\tau)$ is nonstandard and depends on nuisance parameters $(\sigma_w^2, \sigma_{w\psi}(\tau))$ as B_w and B_ψ^τ are correlated Brownian motions.

The limiting distribution of $t_n(\tau)$ can be decomposed as a linear combination of two (independent) distributions, with weights determined by a long-run (zero frequency) correlation coefficient that can be consistently estimated. Following Hansen and Phillips (1990), we have

$$\int_0^1 \underline{B}_w dB_\psi^\tau = \int \underline{B}_w dB_{\psi.w}^\tau + \lambda_{\omega\psi}(\tau) \int \underline{B}_w dB_w,$$

where $\lambda_{\omega\psi}(\tau) = \sigma_{w\psi}(\tau)/\sigma_w^2$ and $B_{\psi.w}^\tau$ is a Brownian motion with variance $\sigma_{\psi.w}^2(\tau) = \sigma_\psi^2(\tau) - \sigma_{w\psi}^2(\tau)/\sigma_w^2$ and is independent of $\underline{B}_w$. Therefore, the limiting distribution of $t_n(\tau)$ can be decomposed as

$$\frac{1}{\sqrt{\tau(1-\tau)}} \frac{\int \underline{B}_w dB_{\psi.w}^\tau}{\left(\int_0^1 \underline{B}_w^2\right)^{1/2}} + \frac{\lambda_{w\psi}(\tau)}{\sqrt{\tau(1-\tau)}} \frac{\int \underline{B}_w dB_w}{\left(\int_0^1 \underline{B}_w^2\right)^{1/2}}.$$

For convenience of exposition, we may rewrite the Brownian motions $B_w(r)$ and $B_{\psi.w}^\tau(r)$ as

$$B_w(r) = \sigma_w W_1(r), \; B_{\psi.w}^\tau(r) = \sigma_{\psi.w}(\tau) W_2(r),$$

$$\underline{B}_w(r) = \sigma_w \underline{W}_1(r), \; \underline{W}_1(r) = W_1(r) - \int_0^1 W_1(s) ds,$$

where $W_1(r)$ and $W_2(r)$ are standard Brownian motions and are independent of one another. Note that $\sigma_\psi^2(\tau) = \tau(1-\tau)$, and the limiting distribution of $t_n(\tau)$ can be written as,

$$\delta \left(\int_0^1 \underline{W}_1{}^2 \right)^{-1/2} \int_0^1 \underline{W}_1 dW_1 + \sqrt{1-\delta^2} N(0,1), \tag{44}$$

where

$$\delta = \delta(\tau) = \frac{\sigma_{w\psi}(\tau)}{\sigma_w \sigma_\psi(\tau)} = \frac{\sigma_{w\psi}(\tau)}{\sigma_w \sqrt{\tau(1-\tau)}}.$$

The above limiting distribution can be easily approximated using simulation methods. In fact, required critical values are tabulated in the literature and thus are available for use in applications.

Alternatively, we may consider a transformation of $t_n(\tau)$ that annihilates the nuisance parameter, and thereby provides a distributional-free form of inference. Hasan and Koenker (1997) consider rank-type tests based on regression rank scores in an augmented Dickey-Fuller framework. A third option is to abandon the asymptotically distribution-free nature of tests and use critical values generated by resampling methods. One may also consider unit root tests based on quantile autoregression over a range of (multiple) quantiles – targeting toward a somewhat broader class of alternatives than those considered in the OLS literature. See Koenker and Xiao (2004) for more discussions on this topic.

8.2. *Quantile regression on cointegrated time series*

Consider again the regression model (27), if X_t is a k-dimensional vector of integrated regressors and u_t is still mean zero stationary (possibly correlated with X_t), it becomes the important cointegration regression model (Xiao, 2009). To deal with endogeneity, we may use leads and lags of X_t (other methods may also be considered to deal with correlation between u_t and X_t). If we assume that u_t has the following representation

$$u_t = \sum_{j=-K}^{K} v_{t-j}' \Pi_j + \varepsilon_t, \tag{45}$$

where $v_t = \Delta X_t$, ε_t is a stationary process, such that $E(v_{t-j}\varepsilon_t) = 0$, for any j, and

$$n^{-1/2} \sum_{t=1}^{[nr]} \begin{bmatrix} \psi_\tau(\varepsilon_{t\tau}) \\ v_t \end{bmatrix} \Rightarrow B(r) = \begin{bmatrix} B_\psi^*(r) \\ B_v(r) \end{bmatrix} = BM(0, \Omega^*),$$

the original cointegrating regression can be rewritten as:

$$Y_t = \alpha + \beta' X_t + \sum_{j=-K}^{K} \Delta X_{t-j}' \Pi_j + \varepsilon_t.$$

If we denote the τth quantile of ε_t as $Q_\varepsilon(\tau)$, let $\mathcal{G}_t = \sigma\{X_t, \Delta X_{t-j}, \forall j\}$, then, conditional on $\mathcal{G}_t$, the τth quantile of Y_t is given by

$$Q_{Y_t}(\tau|\mathcal{G}_t) = \alpha + \beta' X_t + \sum_{j=-K}^{K} \Delta X'_{t-j}\Pi_j + F_\varepsilon^{-1}(\tau),$$

where $F_\varepsilon(\cdot)$ is the c.d.f. of ε_t. Let Z_t be the vector of regressors consisting $z_t = (1, X_t)$ and $(\Delta X'_{t-j}, j = -K, \ldots, K)$, $\Theta = (\alpha, \beta', \Pi'_{-K}, \ldots, \Pi'_K)'$, and

$$\Theta(\tau) = (\alpha(\tau), \beta(\tau)', \Pi'_{-K}, \ldots, \Pi'_K)',$$

where $\alpha(\tau) = \alpha + F_\varepsilon^{-1}(\tau)$, then, we can rewrite the above regression as $Y_t = \Theta' Z_t + \varepsilon_t$, and

$$Q_{Y_t}(\tau|\mathcal{F}_t) = \Theta(\tau)' Z_t.$$

We now consider the following quantile cointegrating regression:

$$\widehat{\Theta}(\tau) = \arg\min_\theta \sum_{t=1}^{n} \rho_\tau(Y_t - \Theta' Z_t). \tag{46}$$

Similar to case of the ADF regression, the components in $\widehat{\Theta}(\tau)$ have different rates of convergence. Denote $G_n = diag(\sqrt{n}, n, \ldots, n, \sqrt{n}, \ldots, \sqrt{n})$. Conformable with $\Theta(\tau)$, we partition $\widehat{\Theta}(\tau)$ as follows:

$$\widehat{\Theta}(\tau)' = \left[\widehat{\alpha}(\tau), \widehat{\beta}(\tau)', \widehat{\Pi}_{-K}(\tau)', \ldots, \widehat{\Pi}_K(\tau)'\right].$$

Under regularity assumptions,

$$G_n(\widehat{\Theta}(\tau) - \Theta(\tau)) \Rightarrow \frac{1}{f_\varepsilon(F_\varepsilon^{-1}(\tau))} \begin{bmatrix} \int_0^1 \overline{B}_v \overline{B}_v^\top & 0 \\ 0 & \Gamma \end{bmatrix}^{-1} \begin{bmatrix} \int_0^1 \overline{B}_v dB_\psi^* \\ \Psi \end{bmatrix}.$$

In particular,

$$n(\widehat{\beta}(\tau) - \beta(\tau)) \Rightarrow \frac{1}{f_\varepsilon(F_\varepsilon^{-1}(\tau))} \left[\int_0^1 \underline{B}_v \underline{B}_v^\top\right]^{-1} \int_0^1 \underline{B}_v dB_\psi^*,$$

and where $\overline{B}_v(r) = (1, B_v(r)')'$, and $\underline{B}_v(r) = B_v(r) - rB_v(1)$, $\Gamma = E(V_t V_t')$, and $V_t = (\Delta X'_{t-K}, \ldots, \Delta X'_{t+K})'$, and Ψ is a multivariate normal with dimension conformable with $(\Pi_{-K}(\tau)', \ldots, \Pi_K(\tau)')'$.

Consider the quantile regression residual

$$\varepsilon_{t\tau} = Y_t - Q_{Y_t}(\tau|\mathcal{F}_t) = Y_t - \Theta(\tau)' Z_t = \varepsilon_t - F_\varepsilon^{-1}(\tau),$$

then we have $Q_{\varepsilon_{t\tau}}(\tau) = 0$, where $Q_{\varepsilon_{t\tau}}(\tau)$ signifies the τth quantile of $\varepsilon_{t\tau}$, and $E\psi_\tau(\varepsilon_{t\tau}) = 0$.

The cointegration relationship may be tested by directly looking at the fluctuation in the residual process $\varepsilon_{t\tau}$ from the quantile cointegrating regression. If we consider the following partial sum process

$$Y_n(r) = \frac{1}{\omega_\psi^* \sqrt{n}} \sum_{j=1}^{[nr]} \psi_\tau(\varepsilon_{j\tau}),$$

where ω_ψ^{*2} is the long-run variance of $\psi_\tau(\varepsilon_{j\tau})$, under appropriate assumptions, the partial sum process follows an invariance principle and converges weakly to a standard Brownian motion $W(r)$. Choosing a continuous functional $h(\cdot)$ that measures the fluctuation of $Y_n(r)$, notice that $\psi_\tau(\varepsilon_{j\tau})$ is indicator based, a robust test for cointegration can be constructed based on $h(Y_n(r))$. By the continuous mapping theorem, under regularity conditions and the null of cointegration,

$$h(Y_n(r)) \Rightarrow h(W(r)).$$

In principle, any metric that measures the fluctuation in $Y_n(r)$ is a natural candidate for the functional h. The classical Kolmogoroff–Smirnof-type or Cramer–von Mises-type measures are of particular interest. Under the alternative of no cointegration, the statistic diverges.

In practice, we estimate $\Theta(\tau)$ by $\widehat{\Theta}(\tau)$ using (46) and obtain the residuals

$$\widehat{\varepsilon}_{t\tau} = Y_t - \widehat{\Theta}(\tau)' Z_t.$$

A robust test for cointegration can then be constructed based on

$$\widehat{Y}_n(r) = \frac{1}{\widehat{\omega}_\psi^* \sqrt{n}} \sum_{j=1}^{[nr]} \psi_\tau(\widehat{\varepsilon}_{j\tau}),$$

where $\widehat{\omega}_\psi^{*2}$ is a consistent estimator of ω_ψ^{*2}. Under regularity assumptions and the hypothesis of cointegration,

$$\widehat{Y}_n(r) \Rightarrow \widetilde{W}(r) = W_1(r) - \left[\int_0^1 dW_1 \overline{W}_2'\right] \left[\int_0^1 \overline{W}_2 \overline{W}_2'\right]^{-1} \int_0^r \overline{W}_2(s),$$

where $\overline{W}_2(r) = (1, W_2(r)')'$ and W_1 and W_2 are independent 1 and k-dimensional standard Brownian motions – see Xiao (forthcoming) for more discussion on robust tests for cointegration.

9. Time series quantile regression applications

There is a large and growing literature of quantile regression applications in various fields. We discuss three examples of quantile regression applications in this section: interval forecasting, testing for structural changes, and portfolio construction.

9.1. Forecasting with quantile models

Dynamic quantile regression models offer a natural approach for interval forecasting. Given the time series observations $\{Y_t\}_{t=1}^T$, and a dynamic quantile regression model

$$Q_{Y_t}(\tau|\mathcal{F}_{t-1}) = g(X_t, \theta(\tau)),$$

where $X_t = (1, Y_{t-1}, \ldots, Y_{t-p})^\top$, we consider out-of-sample prediction based on the available observations. In the special case $g(X_t, \theta(\tau)) = X_t^\top \theta(\tau)$, this is the QAR model (5). If the parameters $\theta(\tau)$ were known, the interval

$$[g(X_{T+1}, \theta(\alpha/2)), g(X_{T+1}, \theta(1-\alpha/2))]$$

is an exact $1-\alpha$ level interval forecast of Y_{T+1}. In practice, we do not know $\theta(\tau)$ and have to use a quantile regression estimator $\widehat{\theta}(\tau)$ in the above construction. Following Portnoy and Zhou (1996), we may use the following modified interval forecast

$$\left[g(X_{T+1}, \widehat{\theta}(\alpha/2 - h_T)), g(X_{T+1}, \widehat{\theta}(1-\alpha/2 + h_T))\right],$$

where $h_T \to 0$, to account for the uncertainty from the preliminary quantile regression estimation $\widehat{\theta}(\tau)$.

To generate an p-step interval forecast for Y_{T+p}, notice that the one-step ahead forecast conditional distribution of Y_{T+1} can be obtained from

$$\widehat{Y}_{T+1} = g(X_{T+1}, \widehat{\theta}(U)),$$

where U are random draws from a uniform distribution $U[0,1]$, let U_1^* be a draw from uniformly distribution on [0,1], then a draw from the one-step ahead forecast distribution of Y_{T+1} is given by

$$\widehat{Y}_{T+1}^* = g(X_{T+1}, \widehat{\theta}(U_1^*)).$$

Next, let $\widetilde{X}_{T+2} = (1, \widehat{Y}_{T+1}^*, Y_T, \ldots, Y_{T-p+2})^\top$, and $U_2^* \sim U[0,1]$, then a draw from the two-step ahead forecast distribution of Y_{T+2} is given by

$$\widehat{Y}_{T+2}^* = g(\widetilde{X}_{T+2}, \widehat{\theta}(U_2^*)).$$

At step s, let $\widetilde{X}_{T+s} = (1, \widehat{Y}_{T+s-1}^*, \ldots, \widehat{Y}_{T+s-p}^*)^\top$ (where $\widehat{Y}_j^* = Y_j$ if $j \le T$) and $U_s^* \sim U[0,1]$, we can obtain a draw of forecast

$$\widehat{Y}_{T+s}^* = g(\widetilde{X}_{T+s}, \widehat{\theta}(U_s^*)).$$

Applying the above sampling procedure recursively, we obtain a sample path of forecast

$$\left(\widehat{Y}_{T+1}^*, \widehat{Y}_{T+2}^*, \ldots, \widehat{Y}_{T+p}^*\right).$$

Repeating this process R times, a forecast of the conditional distribution of Y_{T+p} can then be approximated based on an ensemble of such sample paths $\left\{(\widehat{Y}_{T+1}^{(r)},\ldots,\widehat{Y}_{T+p}^{(r)})\right\}_{r=1}^{R}$, and an p-step $(1-\alpha)$ level interval forecast can be constructed based on the sample quantiles of $\left\{\widehat{Y}_{T+p}^{(r)}\right\}_{r=1}^{R}$.

Other types of models, say, the ARCH model, or a combination of AR structure in the mean equation and ARCH error, may be forecasted in the same way – see Granger et al. (1989) and Koenker and Zhao (1996) for a discussion based on the linear ARCH models.

Yu and Moyeed (2001) proposed a Bayesian solution to the quantile regression problem via the likelihood of a skewed-Laplace distribution. If the density of a random variable takes the form

$$f(\varepsilon)=\tau(1-\tau)\exp\{-\rho_\tau(\varepsilon)\},$$

the density is called an asymmetric Laplace density. Consider, say, an autoregression model, if the error term u_t has probability density function $\propto \exp\{-\rho_\tau(u)\}$, the associated maximum likelihood estimation is equivalent to minimizing the check function of a quantile regression. The likelihood interpretation of quantile regression facilitates extracting the posterior distributions of unknown parameters via the MCMC method and provides a convenient way of incorporating parameter uncertainty into quantile predictive inference. Giving a specified quantile model, Bayesian quantile forecasting can be obtained along this direction.

Lee and Yang (2007) proposed bootstrap aggregating (bagging) to generate quantile predictors.

9.2. *Testing for structural changes in conditional distribution*

Quantile regression offers a variety of techniques for making inferences about conditional quantile functions. For example, it can provide a useful approach in testing for changes in distribution or conditional distribution. Being the inverse of a conditional distribution function, the conditional quantile function is a natural object to examining distributional changes.

Let $\{Y_t,X_t\}_{t=1}^{n}$ denote a time series sequence of random vectors, $\mathcal{F}_0=\sigma\{X_1\}$, $\mathcal{F}_{t-1}=\sigma(Y_{t-1},\ldots,Y_1,X_t,\ldots,X_1)$ for $t\geq 2$, and assume that the τth conditional quantile function of Y_t given $\mathcal{F}_{t-1}$ is given by:

$$Q_{Y_t}(\tau|\mathcal{F}_{t-1})=\beta(\tau,t)'X_t, \tag{47}$$

where $\beta(\tau,t)$ is a $p\times 1$ parameter vector. For example, if $X_t=\left(1,Y_{t-1},\ldots,Y_{t-p-1}\right)'$ and $\beta(\tau,t)$ does not depend on t, we get the quantile autoregression (QAR) model (5),

$$Q_{Y_t}(\tau|\mathcal{F}_{t-1})=\beta_1(\tau)+\beta_2(\tau)Y_{t-1}+\cdots+\beta_p(\tau)Y_{t-p-1}=\beta(\tau)'X_t, \tag{48}$$

where $\beta(\tau)=\left(\beta_1(\tau),\beta_2(\tau),\ldots,\beta_p(\tau)\right)'$.

We are interested in testing the null hypothesis that, in a random sample of size n, the conditional distribution of Y_t, given X_t, has not changed. Let the conditional distribution function be $F_t(y, x_t) = \Pr(Y_t \leq y|x_t)$, then, the null hypothesis can be written as

$$H_0 : F_t(y, x_t) = F(y, x_t).$$

Because the inverse of a conditional distribution function is the conditional quantile function, we can equivalently express H_0 as

$$H_0 : Q_t(\tau, x_t) = Q(\tau, x_t),$$

where $Q_t(\tau, x_t)$ and $Q(\tau, x_t)$ are conditional quantile functions of Y_t, given x_t, obtained from solving

$$F_t(y, x_t) = \tau \text{ and } F(y, x_t) = \tau, \text{ respectively, for } \tau \in [0, 1].$$

If we consider a linear parametric model, such that

$$Q_t(\tau, x_t) = \beta(\tau, t)' x_t,$$

we can then write the inference problem in terms

$$H_0 : \beta(\tau, t) = \beta(\tau) \text{ for some } \beta(\tau) \in \mathcal{B} \subset \mathbb{R}^p. \tag{49}$$

If there is a change in distribution, as the change point r is usually unknown in practice, we have to endogenize it. For this purpose, we define a dummy variable $I_{r,t} = 1(t \geq \lceil nr \rfloor + 1)$, where $1(\cdot)$ is the indicator function. We consider the sequential quantile regression model

$$Q_{Y_t}\left(\tau|x_t, I_{r,t}\right) = \beta(\tau)' x_t + \delta(\tau)' \left(x_t I_{r,t}\right), \tag{50}$$

then testing the null of no structural change reduces to testing

$$H_0 : \delta_0(\tau) = 0 \text{ for all } \tau, \tag{51}$$

where $\delta_0(\tau)$ is the true parameter value of $\delta(\tau)$ in (50).

To proceed, let $z_{rt} = \left(x_t', x_t' I_{r,t}\right)'$ and $\theta(\tau) = (\beta(\tau)', \delta(\tau)')'$. Based on $\{Y_t, z_{rt}\}_{t=1}^n$, the sequential quantile regression estimators (SQREs) of $\theta(\tau)$ are given by

$$\widehat{\theta}(\tau, r) = \arg\min_{\beta \in \mathbb{R}^{2p}} \rho_\tau\left(Y_t - \theta(\tau)' z_{rt}\right), \tag{52}$$

where $\rho_\tau(u) = u[\tau - 1(u < 0)]$. $\widehat{\theta}(\tau, r) = (\widehat{\beta}(\tau, r)', \widehat{\delta}(\tau, r)')' \in \mathbb{R}^p \times \mathbb{R}^p$. Intuitively, under the null hypothesis, we expect that $\widehat{\delta}(\tau, r)$ should be small for all τ and r.

Su and Xiao (2008) studied this application. Let $F(\cdot|\mathcal{F}_{t-1})$ denote the conditional distribution function of Y_t given $\mathcal{F}_{t-1}$. $F(\cdot|\mathcal{F}_{t-1}) = F(\cdot|x_t) = F_t(\cdot)$. $F_t(\cdot)$ has

Lebesgue density $f_t(\cdot) = f(\cdot|x_t)$ a.s., assuming that under the null hypothesis of no distributional change, $\mathrm{E}[\psi_\tau(Y_t - \beta_0(\tau)' x_t)|\mathcal{F}_{t-1}] = 0$ a.s. for some unique $\beta_0(\tau) \in \mathcal{B} \subset \mathbb{R}^p$, and $\beta_0(\tau)$ is an interior point of the compact set $\mathcal{B}$ for each τ, under assumptions on the conditional distribution and weak dependence of the time series, it can be shown that the sequential regression quantile process $\sqrt{n}\left(\widehat{\theta}(\tau, r) - \theta_0(\tau)\right)$ have the following Bahadur representation uniformly in both τ and r:

$$\left[\frac{1}{n}\sum_{t=1}^{n} f_t(\theta_0(\tau)' z_{rt}) z_{rt} z_{rt}'\right]^{-1} \frac{1}{\sqrt{n}}\sum_{t=1}^{n} \psi_\tau(Y_t - \theta_0(\tau)' z_{rt}) z_{rt}.$$

If we assume that $\sup_{0<r\leq 1}\left|n^{-1}\sum_{t=1}^{\lceil nr\rfloor} x_t x_t' - rQ\right| = o_P(1)$, where Q is a finite, symmetric, and positive definite matrix, and

$$\sup_{0\leq\tau\leq 1,\ 0<r\leq 1}\left|n^{-1}\sum_{t=1}^{\lceil nr\rfloor} f_t\left(\beta_0(\tau)' x_t\right) x_t x_t' - rH^*(\tau)\right| = o_P(1),$$

where $H^*(\tau)$ is a finite, symmetric, and positive definite matrix for each τ, under the null,

$$\sqrt{n}\,\widehat{\delta}(\tau, r) \Rightarrow (r(1-r))^{-1} H^*(\tau)^{-1} Q^{1/2} W(\tau, r),$$

where $W(\tau, r) = rW^*(\tau, 1) - W^*(\tau, r)$, and $\{W^*(\tau, r) : (\tau, r) \in [0,1]^2\}$ is a Kiefer process with $E[W^*(\tau, r)] = 0$ and $E[W^*(\tau_1, r_1) W^*(\tau_2, r_2)] = (r_1 \wedge r_2)(\tau_1 \wedge \tau_2 - \tau_1\tau_2) I_p$.

Let $\widehat{\Omega}(\tau, r)$ be a uniformly consistent estimator of

$$\Omega(\tau, r) \equiv \frac{\tau(1-\tau)}{r(1-r)} H^*(\tau)^{-1} Q H^*(\tau)^{-1},$$

the following sup-Wald statistic can be used to test for structural changes in conditional distribution:

$$\sup W_n \equiv \sup_{\tau\in\mathcal{T}}\sup_{r\in\mathcal{A}} W_n(\tau, r) \quad \text{with } W_n(\tau, r) = n\widehat{\delta}(\tau, r)' \widehat{\Omega}(\tau, r)^{-1} \widehat{\delta}(\tau, r). \quad (53)$$

Under the null and regularity conditions,

$$\sup W_n \xrightarrow{d} \sup_{\tau\in\mathcal{T}}\sup_{r\in\mathcal{A}} W(\tau, r)' W(\tau, r) / [\tau(1-\tau) r(1-r)].$$

See Su and Xiao (2008) and Qu (2008) for related studies in linear quantile regression models. Also see Hušková (1997) and Hušková and Picek (2002) for related studies.

In the above model, notice that $\mathcal{F}_{t-1}$ contains y_{t-1}; hence, under the assumption

$$\mathrm{E}[\psi_\tau(Y_t - \beta_0(\tau)' X_t)|\mathcal{F}_{t-1}] = 0,$$

the quantile regression residual is a martingale difference sequence. This assumption can be relaxed, and the above analysis can be extended to the case where the regression residuals are weakly correlated. In this case, under the null hypothesis of no structural change in the conditional distribution, the limiting regression quantile process is still a zero-mean Gaussian process. However, the covariance kernel of the limiting process will be more complicate and depends on the dependence structure in the data. Inference procedures can be constructed based on the simulation methods.

In the above case, the relationship between Y_t and X_t is characterized by a parametric model; testing for distributional change may be formulated as testing quantile regression coefficient instability. In many applications, the functional form of the relationship between Y_t and X_t is unknown. Misspecification of econometric models can also manifest themselves in the form of structural changes. Misleading conclusions may be obtained if the linearity (or other parametric) assumption is violated. To avoid spurious breaks from misspecification, Su and Xiao (2009) proposed and studied residual-based tests for distributional changes via nonparametric quantile regressions.

9.3. *Portfolio construction*

There is a large literature of quantile regression applications in finance – see, e.g., Taylor (1999), Chernozhukov and Umantsev (2001), Bassett and Chen (2001), Wu and Xiao (2002), Bassett et al. (2004), Linton and Whang (2004), Ma and Pohlman (2008), Gowlland et al. (2009), and Xiao and Koenker (2009). Quantile regression has very important applications in portfolio construction. There has been an ongoing debate in the financial literature about which risk measures to use in market risk measurement and portfolio selection. For a long time, the variance of portfolio return has been the predominant market risk measurement. Markowitz (1952) proposed that investors should choose the portfolio that offers the smallest return variance for a given level of expected return. This approach to optimal portfolio selection has a nice connection to maximizing expected utility if portfolio returns are normally distributed or if investors have quadratic utility.

However, in general, financial returns are not normally distributed. Empirical evidence against the normality of returns has been reported by many researchers. In empirical analysis, financial time series tend to be heavy tailed (or "leptokurtic"), and these features are usually accentuated when the data are sampled more frequently. On the other hand, quadratic utility assumes that investors are as averse to upside gain as they are to downside loss. In practice, investors care mainly about the loss associated with downside movements, and upside gain should not be penalized. Over the last few decades, accumulated empirical evidence indicates that when people make investment decisions, they often have different attitude with respect to gains and losses.

A growing number of researchers and practitioners are using downside-risk measurements in various portfolio management applications. The most prominent example is Value-at-Risk (VaR)-based risk measurement, which has become a part of the international banking regulatory mechanisms. Value-at-Risk is defined as the *percentage* loss in market value over a given time horizon that is exceeded with probability τ. That is, for a time series of returns on an asset, $\{r_t\}_{t=1}^n$, the Value-at-Risk at time t, VaR_t, is defined by

$$\Pr\left(r_t < -VaR_t | \mathcal{I}_{t-1}\right) = \tau, \tag{54}$$

where $\mathcal{I}_{t-1}$ denotes the information set at time $t-1$. VaR is a conditional quantile, and therefore, estimation of Value-at-Risk is intimately linked to quantile estimation, and the quantile regression models introduced in the previous sections can be applied to these problems.

Minimizing portfolio VaR would imply that investors only care about τ th quantile of portfolio return distribution, instead of paying attention to the whole distribution of portfolio return. Despite of its popularity, VaR as a risk measure has been criticized by financial engineers. An important criticism of VaR is that it is not a "coherent" risk measure. Following the axiomatic approach, Artzner et al. (1999) define a coherent risk measurement from a regulator's point of view. In the mean-variance-based approach, the standard deviation (variance) is used as a measure of risk. According to the definition of Artzner et al. (1999), such a measure is not a coherent risk measurement.

VaR is also not coherent. For this reason, Expected Shortfall (ES), a coherent risk measurement, has been suggested as an alternative (remedy) for VaR-based risk measurement. ES is defined as the expected loss exceeding VaR. More specifically, The ES of portfolio return Y at τ level is the expected loss of portfolio value given that a loss is occurring at or below the τth quantile:

$$\mathrm{ES}_\tau = E(Y|Y < \mathrm{VaR}_\tau).$$

Unlike VaR, which is insensitive to the magnitude of loss beyond a certain percentile, ES weights large losses by their magnitude. Because ES has the nice property of being a coherent risk measurement, researchers and practitioners, recently, advocate the Mean-ES analysis for portfolio selection by minimizing portfolio ES for a given level of expected portfolio return level. This approach is usually called the Mean-ES analysis.

Bassett et al. (2004) studied Mean-ES portfolio allocation and Choquet expected utility maximization via quantile regression. Let q_τ be the τth quantile of the return distribution, the Mean-ES approach corresponds to a simple truncated utility function

$$u(R) = \begin{cases} R/\tau, & \text{if } R \le q_\tau, \\ 0, & \text{otherwise.} \end{cases} \tag{55}$$

In the Mean-ES setting, investors expected that utility is the ES of the portfolio return distribution corresponding to τth quantile:

$$Eu(R) = \int_{-\infty}^{+\infty} u(R)dF(R) = \frac{1}{\tau}\int_{-\infty}^{q_\tau} RdF(R).$$

Bassett et al. (2004) define the τ-risk of R as

$$\varrho_\tau(R) = -\frac{1}{\tau}\int_{-\infty}^{q_\tau} RdF(R),$$

where q_τ is the τth quantile of the return distribution, and show that empirical strategies for minimizing the τ-risk lead to the methods of quantile regression. Let $\rho_\tau(\cdot)$ be the check function of quantile regression, they show that

$$\min_{\theta} \mathrm{E}\rho_\tau(R-\theta) = \alpha\left(\mu + \varrho_\tau(R)\right). \tag{56}$$

Consider an investment decision over L underlying assets with random returns $r = (r_1, \ldots, r_L)'$, if we construct a portfolio by choosing portfolio weights $w = (w_1, \ldots, w_L)'$, $\sum_{i=1}^{L} w_i = 1$, the portfolio return rate R is equal to $w'r$. Denote the mean of the portfolio as $\mu(w'r)$, the optimal portfolio choice for an Mean-ES investor corresponds to

$$\begin{aligned} &\min_{w} \varrho_\tau(w'r) \\ &\text{s.t. } \mu(w'r) = \mu_0, \sum_{i=1}^{L} w_i = 1. \end{aligned}$$

In practice, giving a sample of n observations of the assets return, say $\{r_t, t = 1, \ldots, n\}$, using the relationship (56) and replacing the expectations by their sample analogues, we have

$$\begin{aligned} &\min_{w,\theta} \sum_{t=1}^{n} \rho_\tau(w'r_t - \theta) \\ &\text{s.t. } \frac{1}{n}\sum_{t=1}^{n} w'r_t = \mu_0, \sum_{i=1}^{L} w_i = 1. \end{aligned}$$

To study the above quantile regression problem and incorporate the restriction $\sum_{j=1}^{L} \omega_j = 1$, we may transform the data $Y_t = r_{1t}$, $X_t = (r_{1t} - r_{2t}, \ldots, r_{1t} - r_{Lt})^\top$, and let

$$\omega = \left(1 - \sum_{j=2}^{L} \beta_j, \beta_2, \ldots, \beta_L\right), \beta = (\beta_2, \ldots, \beta_L)^\top,$$

then the problem can be rewritten as the following unrestricted quantile regression

$$\min_{\beta,\theta} \left\{ \sum_{t=1}^{n} \rho_\tau\left(Y_t - \theta - \beta^\top X_t\right) \right\},$$

thus providing a straightforward way to estimate optimal portfolio weights using expected shortfall as a risk criterion.

10. Conclusion

Time series quantile regression is a growing subject – with many interesting issues under current investigation. This survey is only a selected review on dynamic quantile models. There are lots of interesting topics that are not included due to space restriction. In particular, we only focus on introduction of time series quantile regression methods, many interesting inference problems and empirical applications are not discussed (see, e.g., Koenker (2005) and Koenker and Xiao (2002, 2006)). There are several existing programs for quantile regression applications. For example, both parametric and nonparametric quantile regression estimations can be implemented by the function **rq()** and **rqss()** in the package **quantreg** in the computing language **R**, and **SAS** now has a suite of procedures modeled closely on the functionality of the R package quantreg.

Acknowledgment

I thank Roger Koenker, Steve Portnoy, and an anonymous referee for their very helpful comments on earlier versions of this chapter. This project is partially supported by Boston College Research Fund.

References

Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858.

Angrist, J., Chernozhukov, V., Fernandez-Val, I., 2005. Quantile regression under misspecification, with an application to the U.S. wage structure. Econometrica 74(2), 539–563. March 2006.

Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., 1999. Coherent measures of risk. Math. Finance 9, 203–228.

Bassett, G., Chen, H., 2001. Portfolio style: return-based attribution using quantile regression. Empir. Econ. 26, 293–305.

Bassett, G., Koenker, R., Kordas, G., 2004. Pessimistic portfolio allocation and choquet expected utility. J. Financ. Econom. 4, 477–492.

Bouyé, E., Salmon, M., 2008. Dynamic copula quantile regressions and tail area dynamic dependence in forex markets. Manuscript, Financial Econometrics Research Centre, Warwick Business School, UK.

Cade, B., Noon, B., 2003. A gentle introduction to quantile regression for ecologists. Front. Ecol. Environ. 1, 412–420.

Cai, Z., 2002. Regression quantiles for time series. Econom. Theory 18, 169–192.

Cai, Z., Xiao, Z., 2010. Semiparametric quantile regression estimation in dynamic models with partially varying coefficients. J. Econom. 167, 413–425.

Cai, Z., Xu, X., 2009. Nonparametric quantile estimations for dynamic smooth coefficient models. JASA December 1, 2008, 103(484), 1595–1608.

Chen, X., 2006. Large Sample Sieve Estimation of Semi-Nonparametric Models. In: Heckman, J., & Leamer, E. (Eds.), Handbook of Econometrics, vol. 6, Part 2. North Holland, pp. 5549–5632.

Chen, X., Koenker, R., Xiao, Z., 2009. Copula-based nonlinear quantile autoregression. Econom. J. 12, 50–67.

Chen, X., Shen, X., 1998. Sieve extremum estimates for weakly dependent data. Econometrica 66(2), 289–314.

Chernozhukov, V., 2005. Extremal quantile regression. Ann. Stat. 33(2), 806–839.

Chernozhukov, V., Hansen, C., 2008. Instrumental variable quantile regression: a robust inference approach. J. Econom. 142, 379–398.

Chernozhukov, V., Umantsev, L., 2001. Conditional value-at-risk: aspects of modeling and estimation. Empir. Econom. 26(1), 271–292.

Davis, R.A., Knight, K., Liu, J., 1992. M-estimation for autoregressions with infinite variance. Stoch. Processes Appl. 40(1), 145–180.
De Rossi, G., Harvey, A., 2009. Quantiles, expectiles and splines. J. Econom. 152(2), 179–185.
Engle, R., Mangenelli, S., 2004. CAViaR: conditional autoregressive value at risk by regression quantiles. J. Bus. Econ. Stat. 22, 367–381.
Fan, J., Yao, Q., Tong, H., 1996. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Biometrika 83, 189–206.
Galvao, A., 2010. Quantile Regression for Dynamic Panel, preprint.
Gourieroux, Jasiak, 2008. Dynamic quantile models. J. Econom. 147(1), 198–205.
Gowlland, C., Xiao, Z., Zeng, Q., 2009. Beyond the central tendency: quantile regression as a tool in quantitative investing. J. Portf. Manag. 35(3), 106–119.
Granger, C.W.G., White, H., Kamstra, M., 1989. Interval forecasting: an analysis based on ARCH-quantile estimators. J. Econom. 40, 87–96.
Hall, P., Wolff, R.C.L., Yao, Q., 1999. Methods for estimating a conditional distribution function. JASA 94, 154–163.
Hasan, M.N., Koenker, R., 1997. Robust rank tests of the unit root hypothesis. Econometrica 65(1), 133–161.
Hercé, M., 1996. Asymptotic theory od LAD estimation in a unit root process with finite variance errors. Econom. Theory 12, 129–153.
Honda, T., 2000. Nonparametric estimation of a conditional quantile for α-mixing processes. Ann. Inst. Stat. Math. 52(3), 459–470.
Honda, T., 2004. Quantile regression in varying coefficient models. J. Stat. Plann. Infer. 121, 113–125.
Hsu, Y.H., 2010. Applications of quantile regression to estimation and detection of some tail characteristics, Ph.D. Dissertation, University of Illinois.
Hušková, M., 1997. L1-test procedures for detection of change. In: Dodge, Y. (Ed.), L1-Statistics Procedures and Related Topics. Institute of Mathematical Statistics, Hayward, California, pp. 57–70.
Hušková, M., Picek, J., 2002. M-tests for detection of structural changes in regression. In: Dodge, Y. (Ed.), Statistical Data Analysis Based on the L1-Norm and Related Methods. Birkhauser Verlag, Basel, Switzerland, pp. 213–227.
Jureckova, J., Hallin, M., 1999. Optimal tests for autoregressive models based on autoregression rank scores. Ann. Stat. 27, 1385–1414.
Knight, K., 1989. Limit theory for autoregressive-parameter estimates in an infinite-variance random walk. Can. J. Stat. 17, 261–278.
Knight, K., 1997. Some limit theory for L1-estimators in autoregressive models under general conditions, Lecture Notes-Monograph Series, vol. 31. In: Dodge, Y. (Ed.), L1-Statistical Procedures and Related Topics. California, pp. 315–328.
Knight, K., 1998. Asymptotics for L1 regression estimates under general conditions. Ann. Stat. 26, 755–770.
Knight, K., 2001. Limiting distributions of linear programming estimators. Extremes 4, 87–103.
Knight, K., 2006. Comment on: quantile autoregression. JASA 101(475), 991–1001.
Koenker, R., 2000. Galton, Edgeworth, Frisch and prospects for quantile regression in econometrics. J. Econom. 95, 347–374.
Koenker, R., 2005. Quantile Regression, Econometric Society Monographs (No. 38). Cambridge University Press, New York.
Koenker, R., 2006. Slides in "Econometrics in Rio" . http://www.econ.uiuc.edu/roger/research/qar/Rio.pdf.
Koenker, R., Bassett, G., 1978. Regression Quantiles. Econometrica V46, 33–49.
Koenker, R., Hallock, K., 2001. Quantile regression. J. Econ. Perspect. 15, 143–156.
Koenker, R., Ng, P., Portnoy, S., 1994. Quantile smoothing splines. Biometrika 81, 673–680.
Koenker, R., Xiao, Z., 2002. Inference on the Quantile Regression Processes. Econometrica 70, 1583–1612.
Koenker, R., Xiao, Z., 2004. Unit root quantile regression inference. JASA 99(467), 775–787.
Koenker, R., Xiao, Z., 2006. Quantile autoregression. JASA 101(475), 980–1006.
Koenker, R., Zhao, Q., 1996. Conditional quantile estimation and inference for ARCH models. Econom. Theory 12, 793–813.
Komunjer, I., Vuong, Q., 2010. Efficient estimation in dynamic conditional quantile models. J. Econom. 157(2), 272–285.
Koul, H., Mukherjee, K., 1994. Regression quantiles and related processes under long range dependent errors. J. Multivar. Anal. 51, 318–317.

Koul, H., Saleh, A.K., 1995. Autoregression quantiles and related rank-scores processes. Ann. Stat. 23(2), 670–689.
Kuan, C.M., 2007. An introduction to quantile regression, preprint.
Lee, T.-H., Yang, Y., 2007. Bagging binary and quantile predictors for time series. J. Econom. 135, 465–497.
Linton, O., Shang, D., 2010. Efficient Estimation of Conditional Risk Measures in a Semiparametric GARCH Model, preprint.
Linton, O., Whang, Y.-J., 2004. A Quantilogram Approach to Evaluating Directional Predictability, preprint.
Ma, L., Pohlman, L., 2008. Return forecasts and optimal portfolio construction: a quantile regression approach. Eur. J. Financ. 14, 409–425.
Markowitz, H., 1952. Portfolio selection. J. Finance 7(1), 77–91.
Portnoy, S., 1991. Asymptotic behavior of regression quantiles in non-stationary, dependent cases. J. Multivar. Anal. 38(1), 100–113.
Portnoy, S., Jureckova, J., 1999. On extreme regression quantiles. Extreme 2, 227–243.
Portnoy, S.L., Zhou, K.Q., 1996. Direct use of regression quantiles to construct confidence sets in linear models. Ann. Stat. 24(1), 287–306.
Qu, Z., 2008. Testing for structural change in regression quantiles. J. Econom. 148, 170–184.
Sim, N., 2009. Modeling Quantile Dependence, Ph.D. Dissertation, Boston College, Massachusetts.
Sim, N., Xiao, Z., 2009. Modeling Quantile Dependence: Estimating the Correlations of International Stock Returns, Working Paper, Boston College, Massachusetts.
Su, L., Xiao, Z., 2008. Testing structural change via regression quantiles. Stat. Probab. Lett. 78(16), 2768–2775.
Su, L., Xiao, Z., 2009. Testing for Structural Change in Conditional Distributions via Quantile Regression. Working paper, Boston College.
Taylor, J., 1999. A quantile regression approach to estimating the distribution of multiperiod returns. J. Deriv. 7, 64–78.
Wei, Y., Pere, A., Koenker, R., He, X., 2006. Quantile regression methods for reference growth curves. Stat. Med. 25, 1369–1382.
Weiss, A., 1991. Estimating nonlinear dynamic models using least absolute error estimation. Econom. Theory 7, 46–68.
Wu, G., Xiao, Z., 2002. An analysis of risk measures. J. Risk 4(4), 53–75.
Xiao, Z., forthcoming. Robust inference in nonstationary time series models. J. Econom.
Xiao, Z., 2009. Quantile cointegrating regression. J. Econom. 150(2), 248–260.
Xiao, Z., Koenker, R., 2009. Conditional quantile estimation and inference for GARCH models. JASA 104 (488), 1696–1712.
Yu, K., Jones, M.C.,1998. Local linear quantile regression. JASA 93, 228–237.
Yu, K., Moyeed, R.A., 2001. Bayesian quantile regression. Stat. Probab. Lett. 54(4), 437–447.
Yu, Lu, Stander, 2003. Quantile regression: applications and current research areas. JRSS(D) Statistician 52(3), 331–350.

This page intentionally left blank

Part V: Biostatistical Applications

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00010-7

Frequency Domain Techniques in the Analysis of DNA Sequences

David S. Stoffer
Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, USA

Abstract

The concept of the spectral envelope for analyzing periodicities in categorical-valued time series was introduced in the statistics literature (Stoffer et al., 1993a) as a computationally simple and general statistical methodology for the harmonic analysis and scaling of non-numeric sequences. In the process of developing the technology, many possible interesting adaptations became apparent; for example, Stoffer and Tyler (1998) consider the maximal squared coherency between two categorical-valued time series. One of the most interesting directions was the use of the technology in the analysis of long DNA sequences. A benefit of the techniques was that it combined rigorous statistical analysis with modern computer power to quickly search for diagnostic patterns within long DNA sequences. The methodology is closely related to frequency domain principal component analysis and canonical correlation analysis of time series, and consequently, these topics are described and summarized in the appendix. In addition to presenting the theory and methods of the spectral envelope and related techniques, various analyses of DNA sequences are included. The investigations focus primarily, but not exclusively, on the analysis of viruses. The problems addressed concern about period lengths in nucleosome positioning signals, optimal alphabets in codon usage, and sequence alignment.

Keywords: spectral analysis, molecular biology, spectral envelope, coherency envelope, categorical time series.

1. Introduction

Rapid accumulation of genomic sequences has increased demand for methods to decipher the genetic information gathered in data banks such as GenBank in the

United States, the DNA Data Bank of Japan (DDBJ), and the European Molecular Biology Laboratory (EMBL). Although many methods have been developed for a thorough microanalysis of short sequences, there is a shortage of powerful procedures for the macroanalyses of long DNA sequences. Combining statistical analysis with modern computer power makes it feasible to search, at high speeds, for diagnostic patterns within long sequences. This combination provides an automated approach to evaluating similarities and differences among patterns in long sequences and aids in the discovery of the biochemical information hidden in these organic molecules.

Is a DNA strand a time series? Briefly, a DNA strand can be viewed as a long string of linked nucleotides. Each nucleotide is composed of a nitrogenous base, a five carbon sugar, and a phosphate group. There are four different bases that can be grouped by size, the pyrimidines, thymine (T) and cytosine (C), and the purines, adenine (A) and guanine (G). The nucleotides are linked together by a backbone of alternating sugar and phosphate groups with the $5'$ carbon of one sugar linked to the $3'$ carbon of the next, giving the string direction. DNA molecules occur naturally as a double helix composed of polynucleotide strands with the bases facing inward. The two strands are complementary, so it is sufficient to represent a DNA molecule by a sequence of bases on a single strand; refer to Fig. 1. Thus, a strand of DNA can be represented as a sequence $\{X_t; t = 1, \ldots, n\}$ of letters, termed base pairs (*bp*), from the finite alphabet {A, C, G, T}.[1] The order of the nucleotides contains the genetic information specific to the organism. Expression of information stored in these molecules is a complex multistage process. One important task is to translate the information stored in the protein-coding sequences (CDS) of the DNA. A common problem in analyzing long DNA sequence data is in identifying CDS that are dispersed throughout the sequence and separated by regions of noncoding (which makes up most of the DNA). Another problem of interest that we will address here is that of matching two DNA sequences, say X_{1t} and X_{2t}. The background behind the problem is discussed in detail in the study by Waterman and Vingron (1994). For example, every new DNA or protein sequence is compared with one or more sequence databases to find similar or homologous sequences that have already been studied, and there are numerous examples of important discoveries resulting from these database searches.

Great effort has been focused on questions about the mechanisms placing and removing nucleosomes along the DNA molecule. The exact location of a nucleosome relative to the DNA sequence can be crucial to the regulatory activity. Accordingly, the nucleosome positioning problem became an early concern of molecular genetics (Komberg, 1974). Studies suggest that a large fraction on most genomes are organized

[1] It is worthwhile to review the allocation of symbols used in nucleotide sequences. Aside from the guanine, adenine, thymine, cytosine (G, A, T, C) alphabet, we have R: purine (adenine or guanine); Y: pyrimidine (thymine or cytosine); W: adenine or thymine (for the *weak* hydrogen bonding interaction between the base pairs); S: guanine or cytosine (for the *strong* hydrogen bonding interaction between the base pairs); M: adenine or cytosine (from aMino); K: guanine or thymine (both have Keto groups in similar positions); H: adenine or thymine or cytosine (or not-G); B: guanine or cytosine or thymine (or not-A); V: guanine or adenine or cytosine (or not-T); D: guanine or adenine or thymine (or not-C); N: guanine or adenine or thymine or cytosine (aNy or uNspecified); X: unknown.

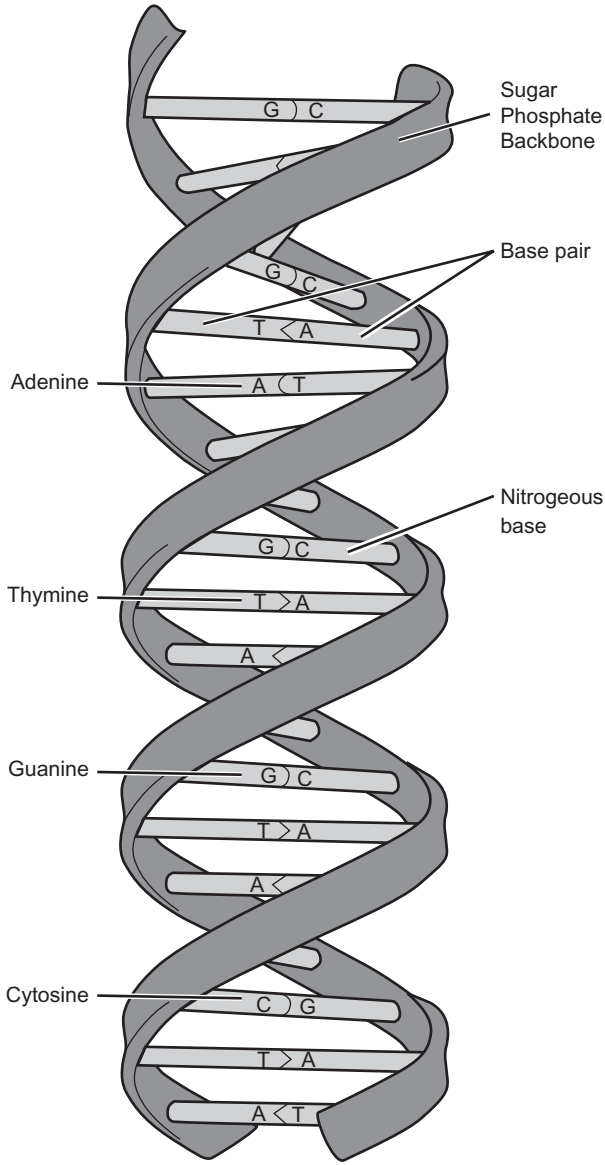

Fig. 1. The general structure of DNA and its bases.

in positioned nucleosomes (for a review, see Simpson (1990)). If positioned nucleosomes do occur in vivo, how are their precise locations established and maintained? Several possible mechanisms have been discussed in the literature, some are strongly supported by empirical evidence, others are not. No evidence, for example, has yet been gathered for special phasing proteins or for replication related alignment. A major factor that seems to contribute to nucleosome positioning is the DNA sequence itself.

Histone–DNA interactions[2] are believed to be decisive in the fine tuning of the precise octamer location. Two parameters are distinguished with respect to sequence-related nucleosome positioning: translational signals mark the site on the DNA sequence, and rotational signals define the curvature of the DNA on the side that faces toward the histones. Translational signals are less well understood at present. One hypothesis suggests that certain preferred base patterns at the dyad[3] serve as translational signals, since special sequence properties are necessary near dyads, where the wrapped DNA deviates in sharp bends from its otherwise smooth superhelical path around the octamer. This view is supported by observed preferences for the trinucleotides RRR-YYY and RRY-RYY in the dyad (Turnell et al., 1988). Another hypothesis regards runs of 5–18 As or Ts as potential signals because they tend to be excluded from the regions near the dyad but are found at both ends. A third sequence-dependent translational signal might be the change, or interruption of periodicities of di- or trinucleotides in the immediate dyad region (Satchwell et al., 1986).

The idea of rotational signals for nucleosome positioning is based on the fact that the nucleosomal DNA is tightly wrapped around its protein core. The bending of the wound DNA requires compression of the grooves that face toward the core and a corresponding widening of the grooves facing the outside. Since, depending on the nucleotide sequence, DNA bends more easily in one plane than another, Trifonov and Sussman (1980) proposed that the association between the DNA sequence and its preferred bending direction might facilitate the necessary folding around the core particle. This sequence-dependent bendability motivated the theoretical and experimental search for rotational signals. These signals were expected to exhibit some kind of periodicity in the sequence, reflecting the structural periodicity of the wound nucleosomal DNA.

Although model calculations as well as experimental data strongly agree that some kind of periodic signal exists, they largely disagree about the exact type of periodicity. A number of questions remain unresolved: Do the periodicities in rotational signals occur predominantly in di- or in trinucleotides, or even in higher order dinucleotides? Ioshikhes et al. (1992) reported new evidence for dinucleotide signals, whereas the analysis of Satchwell et al. (1986) resulted in a trinucleotide pattern that was newly supported by data from the works of Muyldermans and Travers (1994). Which nucleotide alphabets are involved in rotational signals? Satchwell et al. (1986) used a strong (G, C) versus weak (A, T) hydrogen bonding alphabet to propose a 10-bp signal, $W_3N_2S_3N_2$. Zhurkin (1985) suggested the purine–pyrimidine alphabet with an RYN_3YRN_3 pattern, and Trifonov and coworkers propose an AAN_3TTN_3 motif. What is the exact period length? The helical repeat of free DNA is about 10.5 bp, the periodicities of rotational signals tend to be slightly shorter than 10.5 in general, for example: 10.1 bp in Shrader and Crothers (1990), 140.2 in Satchwell et al. (1986), 10.3 bp in Bina (1994), and 10.4 bp in Ioshikhes et al. (1992). Consistent with all these data is the proposition by Shrader and Crothers (1990) that nucleosomal DNA is over wound by about 0.3 bp

[2] Histones are proteins that act as spools around which DNA winds, and play a role in gene regulation. Bending occurs at an approximate period of 10 bp.

[3] Dyad is a type of symmetry that refers to two areas of a DNA molecule whose base pair sequences are inverted relative to each other. The complementary sequences will fold and base-pair with each other, and the sequence of bases between them form a hairpin loop.

per turn. Are there other periodicities besides the approximate 10-bp period? Uberbacher et al. (1988) observed several additional periodic patterns of lengths 6 to 7, 10, and 21 bp. Bina (1994) reports a TT-period of 6.4 bp.

Of course, one could extend this list of controversial questions about the properties and characteristics of positioning signals. Depending on the choice among these divergent observations and claims, different sequence-directed algorithms for nucleosomic mapping have been developed, for example, by Drew and Calladine (1987), Mengeritsky and Trifonov (1983), Uberbacher et al. (1988), Zhurkin (1983), and Piña et al. (1990). The analysis of existing data by the *spectral envelope* (Stoffer et al., 1993a) has resulted in a more unified picture about the major periodic signals that contribute to nucleosome positioning. This, in turn, can lead to new reliable and efficient ways to predict nucleosome locations in long DNA sequences by computer.

In addition to positioning, the spectral envelope has proved to be a useful tool in examining nonsynonymous codon usage. Regional fluctuations in G + C content (isochores) not only influence silent sites, but seem to create a general tendency in high G + C regions toward G + C-rich codons (G + C pressure), see Bernardi and Bernardi (1985) and Sueoka (1988). Schachtel et al. (1991) compared two closely related α-herpesviruses, HSVI, and VZV, and showed that for pairs of homologous genes, G + C frequencies differed in all three codon positions, reflecting the large difference in their global G + C content. In perfect agreement with their overall compositional bias, the usage for each individual amino acid type was shifted significantly toward codons of preferred G + C content. Several authors reported codon context related biases (see Buckingham (1990) for a review). Blaisdell (1983) observed that codon sites three were chosen to be unlike neighboring bases to the left and to the right with respect to the S-W alphabet. Shepherd (1984) observed an enrichment of RNY codons in coding sequences and suggested that this bias was the remnant of a primitive primeval message (see Wong and Cedergren (1986)). Another purine–pyrimidine pattern for weakly expressed genes was suggested by Yarus and Folley (1985). They observed a preference for R|YYR or Y|RRY (the first letter represents the third position of the preceding codon and the bar indicates the border between the codons). Trifonov (1987) and Lagunez-Otero and Trifonov (1992) suggested a G-nonG-N-based frame-keeping mechanism to prevent ribosomal slippage in the translational process. This mechanism could explain a widely observed preference for GHN codons (see Curran and Gross (1994)), for a critical evaluation. Although the various mentioned studies on nonsynonymous codon usage exhibit many substantial differences, most of them agree on one point, namely the existence of some kind of periodicity in coding sequences. This widely accepted observation is supported by the spectral envelope approach that shows a very strong period-three signal in genes but disappears in noncoding regions. This method can even detect wrongly assigned gene segments as will be seen. In addition, the spectral envelope provides not only the optimal period lengths but also most favorable alphabets, for example, {S, W}, {R, Y}, or {G, H}. This analysis might help decide which among the different suggested pattern (such as RNY, GHN, etc.) are the most valid.

The spectral envelope methodology is computationally fast and simple because it is based on the fast Fourier transform and is nonparametric (i.e., it is model independent). This makes the methodology ideal for the analysis of long DNA sequences. Fourier

analysis has been used in the analysis of correlated data (time series) since the turn of the century. Of fundamental interest in the use of Fourier techniques is the discovery of hidden periodicities or regularities in the data. Although Fourier analysis and related signal processing are well established in the physical sciences and engineering, they have only recently been applied in molecular biology. Since a DNA sequence can be regarded as a categorical-valued time series it is of interest to discover ways in which time series methodologies based on Fourier (or spectral) analysis can be applied to discover patterns in a long DNA sequence or similar patterns in two long sequences.

One naive approach for exploring the nature of a DNA sequence is to assign numerical values (or scales) to the nucleotides and then proceed with standard time series methods. It is clear, however, that the analysis will depend on the particular assignment of numerical values. Consider the artificial sequence ACGTACGTACGT... Then, setting A = G = 0 and C = T = 1, yields the numerical sequence 010101010101..., or one cycle every two base pairs (i.e., a frequency of oscillation of $\omega = 1/2$ cycle/bp, or a period of oscillation of length $1/\omega = 2$ bp/cycle). Another interesting scaling is A = 1, C = 2, G = 3, and T = 4, which results in the sequence 123412341234..., or one cycle every four bp ($\omega = 1/4$). In this example, both scalings (i.e., $\{A, C, G, T\} = \{0, 1, 0, 1\}$ and $\{A, C, G, T\} = \{1, 2, 3, 4\}$) of the nucleotides are interesting and bring out different properties of the sequence. It is clear, then, that one does not want to focus on only one scaling. Instead, the focus should be on finding all possible scalings that bring our interesting features of the data. Rather than choose values arbitrarily, the spectral envelope approach selects scales that help emphasize any periodic feature that exists in a DNA sequence of virtually any length in a quick and automated fashion. In addition, the technique can determine whether a sequence is merely a random assignment of letters.

Fourier analysis has been applied successfully in molecular genetics; McLachlan and Stewart (1976) and Eisenberg et al. (1994) studied the periodicity in proteins using Fourier analysis. They used predefined scales (e.g., the hydrophobicity alphabet) and observed the $\omega = 1/3.6$ frequency of amphipathic helices. Because predetermination of the scaling is somewhat arbitrary and may not be optimal, Cornette et al. (1987) reversed the problem and started with a frequency of $\omega_0 = 1/3.6$ and proposed a method to establish an "optimal" scaling at $\omega_0 = 1/3.6$. In this setting, optimality roughly refers to the fact that the scaled (numerical) sequence is maximally correlated with the sinusoid that oscillates at a frequency of ω_0. Viari et al. (1990) generalized this approach to a systematic calculation of a type of spectral envelope (which they called λ-graphs) and of the corresponding optimal scalings over all fundamental frequencies. Although the aforementioned authors dealt exclusively with amino acid sequences, various forms of harmonic analysis have been applied to DNA by, for example, Tavaré and Giddings (1989), and in connection to nucleosome positioning by Satchwell et al. (1986) and Bina (1994). Stoffer et al. (1993a) proposed the spectral envelope as a general technique for analyzing categorical-valued time series in the frequency domain. The basic technique is similar to the methods established by Tavaré and Giddings (1989) and Viari et al. (1990), however, there are some differences. The main difference is that the spectral envelope methodology is developed in a statistical setting to allow the investigator to distinguish between significant results and those results that can be attributed to chance. In particular, tests of significance and confidence intervals can be calculated using large sample techniques.

2. The spectral envelope

2.1. Spectral analysis

For a numerical-valued time series sample, X_t, $t = 1, \ldots, n$, that has been centered by its sample mean, the sample spectral density (or periodogram) is defined in terms of frequency $\omega \in [-\frac{1}{2}, \frac{1}{2}]$ as

$$\widetilde{f}(\omega) = \left| n^{-1/2} \sum_{t=1}^{n} X_t \exp(-2\pi i t \omega) \right|^2.$$

The spectral density $f(\omega)$ of the time series is defined (as a descriptor of the hypothetical population of possible sample paths in the statistical model) as the limit as the sample size n tends to infinity of $E[\widetilde{f}(\omega)]$ provided that it exists. Its existence is guaranteed if the process is stationary with an absolutely summable covariance function, $\gamma(h) = \mathrm{cov}(X_{t+h}, X_t)$; i.e., $\sum_h |\gamma(h)| < \infty$. Details can be found in many time series texts, for example, in the study by. It is worthwhile to note that $f(\omega) \geq 0$, $f(\omega) = f(-\omega)$, and

$$\int_{-1/2}^{1/2} f(\omega)\, d\omega = 2 \int_{0}^{1/2} f(\omega)\, d\omega = \sigma^2 \tag{1}$$

where $\mathrm{var}(X_t) = \sigma^2$ is the population variance of the time series. Thus, the spectral density can be thought of as a decomposition of the total variance of a process into components attributed to frequency. That is, for positive frequencies, the proportion of the variance of X_t that can be attributed to oscillations in the data in the small frequency interval $[\omega, \omega + d\omega]$ is roughly $2f(\omega)d\omega$. If n is a highly composite integer, the fast Fourier transform provides extremely fast calculation of $\widetilde{f}(j/n)$, for $j = 1, 2, \ldots, [\![n/2]\!]$, where $[\![\cdot]\!]$ is the greatest integer function. The frequencies $\omega_j = j/n$ are called the fundamental (or Fourier) frequencies. The sample equivalent of the integral equation (1) is

$$2 \sum_{j=1}^{[\![(n-1)/2]\!]} \widetilde{f}(j/n)\, n^{-1} + \widetilde{f}(1/2)\, n^{-1} = S^2, \tag{2}$$

where S^2 is the sample variance of the data; the last term is dropped if n is odd. One can plot the periodogram, $\widetilde{f}(\omega_j)$, versus the fundamental frequencies $\omega_j = j/n$, for $j = 1, 2, \ldots, [\![n/2]\!]$, and inspects the graph for large values. Large values of the periodogram at ω_j indicate that the data are highly correlated with the sinusoid that is oscillating at a frequency of j cycles in n observations. If the data are uncorrelated (or white noise) the spectral density is flat, that is, $f(\omega) = \sigma^2$ at all frequencies.

Since – no matter how large the sample size – the variance of periodogram is unduly large, the graph of the periodogram can exhibit many nonsignificant peaks. To overcome this problem, a smoothed estimate of the spectral density is typically used.

The general form of the estimate is

$$\widehat{f}(\omega) = \sum_{q=-m}^{m} h_q \widetilde{f}(\omega_{j+q}), \tag{3}$$

where $\{\omega_{j+q}; q = 0, \pm 1, \ldots, \pm m\}$ is a band of frequencies where ω_j is the fundamental frequency closet to ω, and such that the weights $h_q = h_{-q}$ are positive and $\sum_{q=-m}^{m} h_q = 1$. A simple average corresponds to the case where $h_q = 1/(2m+1)$ for $q = -m, \ldots, 0, \ldots, m$. The number m is chosen to obtain a desired degree of smoothness. Larger values of m lead to smoother estimates, but one has to be careful not to smooth away significant peaks (this is the so-called bias-variance tradeoff problem). Experience and trial-and-error can be used to select good values of m and the set of weights $\{h_q\}$. Another consideration is that of tapering the data prior to a spectral analysis; i.e., rather than work with the data X_t directly, one can improve the estimation of spectra by working with tapered data, say $Y_t = a_t X_t$, where tapers $\{a_t\}$ generally have a shape that enhances the center of the data relative to the extremities, such as a cosine bell, $a_t = 0.5[1 + \cos(2\pi t'/n)]$ where $t' = t - (n+1)/2$, favored by Blackman and Tukey (1959). Another related approach is window spectral estimation. Specifically, consider a window function $H(\alpha)$, $-\infty < \alpha < \infty$, that is real-valued, even, of bounded variation, with $\int_{-\infty}^{\infty} H(\alpha)d\alpha = 1$, and $\int_{-\infty}^{\infty} |H(\alpha)|d\alpha < \infty$. The window spectral estimator is

$$\widehat{f}(\omega) = n^{-1} \sum_{q=1}^{n-1} H_n(\omega - q/n)\widetilde{f}(q/n), \tag{4}$$

where $H_n(\alpha) = B_n^{-1} \sum_{j=-\infty}^{\infty} H(B_n^{-1}[\alpha + j])$ and B_n is a bounded sequence of nonnegative scale parameters such that $B_n \to 0$ and $nB_n \to \infty$ as $n \to \infty$. Estimation of the spectral density requires special attention to the issues of leakage and of the variance-bias tradeoff typically associated with the estimation of density functions. Readers who are unfamiliar with this material can consult one of the many texts on the spectral domain analysis of time series; e.g., Shumway and Stoffer (2011), Chapter 4.

An analogous theory applies if one collects p numerical-valued time series, say $X_{1t}, \ldots, X_{pt}$, for $t = 1, \ldots, n$. In this case, write $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{pt})'$ as the $p \times 1$ column vector of data. The periodogram is now a $p \times p$ complex matrix

$$\widetilde{f}(\omega) = \left[n^{-1/2} \sum_{t=1}^{n} \boldsymbol{X}_t \exp(-2\pi i t\omega) \right] \left[n^{-1/2} \sum_{t=1}^{n} \boldsymbol{X}_t \exp(-2\pi i t\omega) \right]^*,$$

where * means to transpose and conjugate. The diagonal elements of $\widetilde{f}(\omega)$ are the individual sample spectra and the off diagonal elements are related to the pairwise dependence structure among the p sequences. We will investigate the off-diagonal elements in more detail later. The population spectral density is again defined as the

limit as n tends to infinity of $E[\widetilde{f}(\omega)]$. Smoothing the periodogram also proceeds analogously to the univariate case, that is, $\widehat{f}(\omega_j) = \sum_{q=-m}^{m} h_q \widetilde{f}(\omega_{j+q})$.

2.2. *Definition and asymptotics*

The spectral envelope is an extension of spectral analysis when the data are categorical-valued such as DNA sequences. To briefly describe the technique using the nucleotide alphabet, let X_t, $t = 1, \ldots, n$ be a DNA sequence taking values in {A, C, G, T}. For real numbers $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)'$, not all equal, denote the scaled (numerical) data by $X_t(\boldsymbol{\beta})$, where

$$X_t(\boldsymbol{\beta}) = \beta_1 \quad \text{if } X_t = \text{A}; \qquad X_t(\boldsymbol{\beta}) = \beta_2 \quad \text{if } X_t = \text{C};$$
$$X_t(\boldsymbol{\beta}) = \beta_3 \quad \text{if } X_t = \text{G}; \qquad X_t(\boldsymbol{\beta}) = \beta_4 \quad \text{if } X_t = \text{T}.$$

For example, if $\boldsymbol{\beta} = (1, 0, 1, 0)'$, then $X_t(\boldsymbol{\beta}) = 1$ if there is a purine (A or G) at position t, and $X_t(\boldsymbol{\beta}) = 0$ if there is a pyrimidine (C or T) at position t. Hence, if X_t is ATAGC, then $X_t(\boldsymbol{\beta})$ is 10110. We define, for each frequency, $\boldsymbol{\beta}(\omega)$ to be the *optimal scaling* at frequency ω if it satisfies

$$\lambda(\omega) = \max_{\boldsymbol{\beta}} \left\{ \frac{f(\omega; \boldsymbol{\beta})}{\sigma_\beta^2} \right\},$$

where $f(\omega; \boldsymbol{\beta})$ is the spectral density of $X_t(\boldsymbol{\beta})$, the scaled data, and σ_β^2 is the variance of the scaled data. Note that $\lambda(\omega)$ can be thought of as the largest proportion of the power (variance) that can be obtained at frequency ω for any scaling of the DNA sequence X_t, and $\boldsymbol{\beta}(\omega)$ is the particular scaling that maximizes the power at frequency ω. Thus, $\lambda(\omega)$ is called the *spectral envelope*. The name spectral envelope is appropriate because $\lambda(\omega)$ envelopes the spectrum of any scaled process. That is, *for any assignment of numbers to letters, the standardized spectral density of a scaled sequence is no bigger than the spectral envelope*, with equality only when the numerical assignment is proportional to the optimal scaling, $\boldsymbol{\beta}(\omega)$. The importance of this fact is demonstrated in Fig. 2. We say "proportional to" because optimal scaling $\boldsymbol{\beta}(\omega)$ is not unique. It is, however, unique up to location and scale changes; that is, any scaling of the form $a\boldsymbol{\beta}(\omega) + b\mathbf{1}$, where $a \neq 0$ and b are real numbers, and $\mathbf{1} = (1, 1, 1, 1)'$ yields the same value of the spectral envelope $\lambda(\omega)$. For example, the numerical assignments {A, C, G, T} = {0, 1, 0, 1} and {A, C, G, T} = {−1, 1, −1, 1} will yield the same normalized spectral density. The value of $\lambda(\omega)$, however, does not depend on the particular choice of scales; details can be found in the works done by Stoffer et al. (1993a). For ease of computation, we set one element of $\boldsymbol{\beta}(\omega)$ equal to zero (i.e., for example, the scale for T is held fixed at T = 0) and then proceed with the computations.

For example, to find the spectral envelope, $\lambda(\omega)$, and the corresponding optimal scaling, $\boldsymbol{\beta}(\omega)$, holding the scale for T fixed at zero, form 3×1 vectors $\boldsymbol{Y}_t$,

$$\boldsymbol{Y}_t = (1, 0, 0)' \quad \text{if } X_t = \text{A}; \qquad \boldsymbol{Y}_t = (0, 1, 0)' \quad \text{if } X_t = \text{C};$$
$$\boldsymbol{Y}_t = (0, 0, 1)' \quad \text{if } X_t = \text{G}; \qquad \boldsymbol{Y}_t = (0, 0, 0)' \quad \text{if } X_t = \text{T}.$$

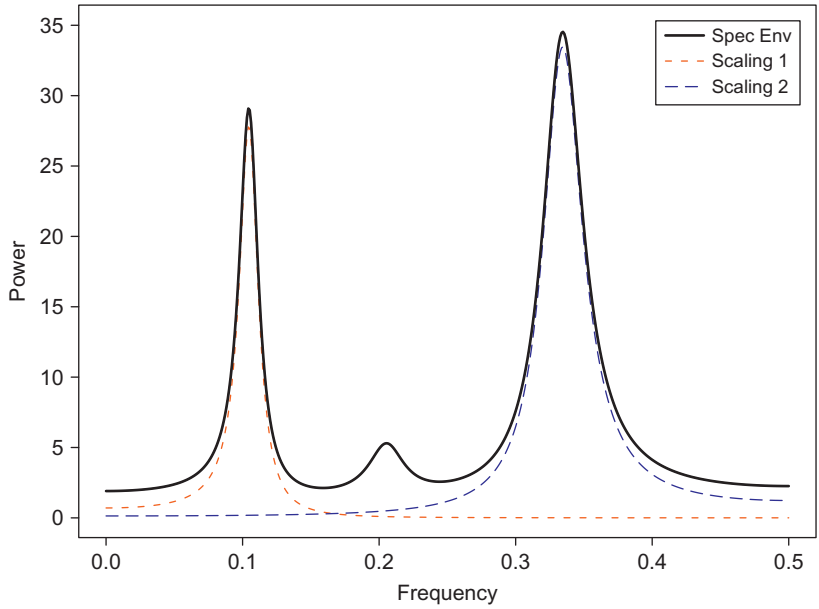

Fig. 2. Demonstration of the spectral envelope. The short dashed line indicates a spectral density corresponding to some scaling. The long dashed line indicates a spectral density corresponding to a different scaling. The thick solid line is the spectral envelope, which can be thought of as throwing a blanket over all possible spectral densities corresponding to all possible scalings of the sequence. Because the exhibited spectral densities attain the value of the spectral envelope at the frequencies near 0.1 and 0.33, the corresponding scalings are optimal at those frequencies. In addition to finding interesting frequencies (e.g., there is something interesting near the frequency of 0.2 that neither scaling 1 or 2 discovers), the spectral envelope reveals frequencies for which nothing is interesting (e.g., no matter which scaling is used, there is nothing interesting in this sequence in the frequency range above 0.4).

Now with $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$, the scaled sequence, $X_t(\boldsymbol{\beta})$, can be obtained from the $\boldsymbol{Y}_t$ vector sequence by the relationship $X_t(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{Y}_t$. This relationship implies that

$$\lambda(\omega) = \max_{\boldsymbol{\beta}} \left\{ \frac{\boldsymbol{\beta}' f_Y(\omega) \boldsymbol{\beta}}{\boldsymbol{\beta}' V \boldsymbol{\beta}} \right\} \tag{5}$$

where $f_Y(\omega)$ is the 3×3 spectral density matrix of the indicator data, $\boldsymbol{Y}_t$, and V is the population variance–covariance matrix of $\boldsymbol{Y}_t$. Because $f_Y(\omega) = f_Y^{re}(\omega) + i\, f_Y^{im}(\omega)$ is Hermitian, $f_Y^{im}(\omega)$ is skew-symmetric, so that $\boldsymbol{\beta}' f_Y(\omega)\boldsymbol{\beta} = \boldsymbol{\beta}' f_Y^{re}(\omega)\boldsymbol{\beta}$. It follows that $\lambda(\omega)$ and $\boldsymbol{\beta}(\omega)$ can easily be obtained by solving an eigenvalue problem with real-valued matrices.

In other words, if $\boldsymbol{Y}_t$ is the vector indicator process associated with a categorical series X_t, and $f_Y(\omega)$ and V are the spectral density and variance–covariance matrices of $\boldsymbol{Y}_t$, respectively, then

(i) The spectral envelope, $\lambda(\omega)$, is the largest eigenvalue of $f_Y^{re}(\omega)$ in the metric of V; that is, $\lambda(\omega)$ is largest eigenvalue of the determinantal equation $\left| f_Y^{re}(\omega) - \lambda V \right| = 0$.
(ii) The optimal scaling $\boldsymbol{\beta}(\omega)$ is the corresponding eigenvector, that is, it satisfies $f_Y^{re}(\omega)\boldsymbol{\beta}(\omega) = \lambda(\omega) V \boldsymbol{\beta}(\omega)$.

An algorithm for estimating the spectral envelope and the optimal scalings given a particular DNA sequence with alphabet $\mathcal{S} = \{c_1, \ldots, c_{k+1}\}$, is as follows.

1. Given a DNA sequence of length n, form the $k \times 1$ vectors $\boldsymbol{Y}_t$, $t = 1, \ldots, n$; namely, for $j = 1, \ldots, k$, $\boldsymbol{Y}_t = \boldsymbol{e}_j$ if $X_t = c_j$ where $\boldsymbol{e}_j$ is a $k \times 1$ vector with a 1 in the jth position as zeros elsewhere, and $\boldsymbol{Y}_t = \boldsymbol{0}$ if $X_t = c_{j+1}$.
2. Calculate the (fast) Fourier transform of the data,

$$\boldsymbol{d}(j/n) = n^{-1/2} \sum_{t=1}^{n} \boldsymbol{Y}_t \exp(-2\pi i t j/n).$$

   Note that $\boldsymbol{d}(j/n)$ is a $k \times 1$ complex-valued vector. Calculate the periodogram, $\widetilde{f}(j/n) = \boldsymbol{d}(j/n)\boldsymbol{d}^*(j/n)$, for $j = 1, \ldots, [\![n/2]\!]$, and retain only the real part, say $f^{re}(j/n)$.
3. Smooth the real part of the periodogram as preferred to obtain $\widehat{f}^{re}(j/n)$, a consistent estimator of the real part of the spectral matrix.
4. Calculate the $k \times k$ variance–covariance matrix of the data, $S = n^{-1} \sum_{t=1}^{n} (\boldsymbol{Y}_t - \overline{\boldsymbol{Y}})(\boldsymbol{Y}_t - \overline{\boldsymbol{Y}})'$, where $\overline{\boldsymbol{Y}}$ is the sample mean of the data.
5. For each $\omega_j = j/n$, $j = 1, \ldots, [\![n/2]\!]$, determine the largest eigenvalue and the corresponding eigenvector of the matrix $2n^{-1} S^{-1/2} \widehat{f}^{re}(\omega_j) S^{-1/2}$. Note that $S^{-1/2}$ is the inverse of the unique square root matrix of S.[4]
6. The sample spectral envelope $\widehat{\lambda}(\omega_j)$ is the eigenvalue obtained in the previous step. If $\boldsymbol{b}(\omega_j)$ denotes the eigenvector obtained in the previous step, the optimal sample scaling is $\widehat{\boldsymbol{\beta}}(\omega_j) = S^{-1/2}\boldsymbol{b}(\omega_j)$; this will result in three values, the fourth being held fixed at zero.

Any standard programming language can be used to do the calculations; basically, one only has to be able to compute fast Fourier transforms and eigenvalues and eigenvectors of real symmetric matrices. Some examples using the R Statistical Programming Language may be found in the works of Shumway and Stoffer (2011, Chapter 7). Again we note that the procedure can be done with any finite number of possible categories, and is not restricted to looking only at the nucleotide alphabets. Inference for the sample spectral envelope and the sample optimal scalings are described in detail by Stoffer et al. (1993a). A few of the main results of that paper are as follows.

If X_t is an i.i.d. sequence, and if no smoothing is used [i.e., $m = 0$ in (3)], then the following large sample approximation based on the chi-square distribution is valid for $x > 0$:

$$\Pr\{n2^{-1}\widehat{\lambda}(\omega_j) < x\} \doteq \Pr\{\chi^2_{2k} < 4x\} - \pi^{1/2} x^{(k-1)/2} \exp(-x) \Pr\{\chi^2_{k+1} < 2x\} \,/\, \Gamma(k/2), \tag{6}$$

where $k + 1$ is the size of the alphabet being considered.

[4] If $S = P\Lambda P'$ is the spectral decomposition of S, then $S^{-1/2} = P\Lambda^{-1/2}P'$, where $\Lambda^{-1/2}$ is the diagonal matrix with the reciprocal of the root eigenvalues along the diagonal.

In the general case, if a smoothed estimator is used and $\lambda(\omega)$ is a distinct root (which implies that $\lambda(\omega) > 0$), then, independently, for any collection of frequencies $\{\omega_{j_i}; i = 1, \ldots, M\}$, M fixed, and for large n and m,

$$\nu_m \frac{\widehat{\lambda}(\omega_{j_i}) - \lambda(\omega_{j_i})}{\lambda(\omega_{j_i})} \sim \text{AN}(0, 1) \tag{7}$$

and

$$\nu_m \left[\widehat{\boldsymbol{\beta}}(\omega_{j_i}) - \boldsymbol{\beta}(\omega_{j_i})\right] \sim \text{AN}(\mathbf{0}, \Sigma_{j_i}), \tag{8}$$

where $\Sigma_{j_i} = V^{-1/2}\Omega_{j_i}V^{-1/2}$ with

$$\Omega_{j_i} = \{\lambda(\omega_{j_i})H(\omega_{j_i})^+ f^{re}(\omega_{j_i})H(\omega_{j_i})^+ - \boldsymbol{a}(\omega_{j_i})\boldsymbol{a}(\omega_{j_i})'\}/2,$$

and $H(\omega_{j_i}) = f^{re}(\omega_{j_i}) - \lambda(\omega_{j_i})\boldsymbol{I}_{k-1}$, $\boldsymbol{a}(\omega_{j_i}) = H(\omega_{j_i})^+ f^{im}(\omega_{j_i})V^{1/2}\boldsymbol{u}(\omega_{j_i})$, and $H(\omega_{j_i})^+$ refers to the Moore–Penrose inverse of $H(\omega_{j_i})$. The term ν_m depends on the type of estimator being used. In the case of weighted averaging, $\nu_m^{-2} = \sum_{q=-m}^{m} h_q^2$ [if a simple average is used, $h_q = 1/(2m + 1)$, then $\nu_m^2 = (2m + 1)$]. Based on these results, asymptotic normal confidence intervals and tests for $\lambda(\omega)$ can be readily constructed. Similarly, for $\boldsymbol{\beta}(\omega)$, asymptotic confidence ellipsoids and chi-square tests can be constructed; details can be found in the study by Stoffer et al. (1993a, Theorems 3.1–3.3). As a note, we mention that this technique is not restricted to the use of sinusoids. In the works done by Stoffer et al. (1993b), the use of the Walsh basis [5] of square-waves functions that take only the values ± 1, is described.

A simple asymptotic test statistic for $\boldsymbol{\beta}(\omega)$ can be obtained. Let $\widehat{H}(\omega) = \widehat{f}_Y^{re}(\omega) - \widehat{\lambda}(\omega)I_k$, and

$$\boldsymbol{\xi}_m(\omega) = \sqrt{2}\,\nu_m\, \widehat{f}_Y^{re}(\omega)^{-1/2}\widehat{H}(\omega)\left(\widehat{\boldsymbol{\beta}}(\omega) - \boldsymbol{\beta}(\omega)\right) \Big/ \widehat{\lambda}(\omega)^{1/2}.$$

Then,

$$\boldsymbol{\xi}_m(\omega)'\boldsymbol{\xi}_m(\omega) \tag{9}$$

converges ($m \to \infty$) in distribution to a distribution that is stochastically less than χ_k^2 and stochastically greater than χ_{k-1}^2. Note that the test statistic (9) is zero if $\boldsymbol{\beta}(\omega)$ is replaced by $\widehat{\boldsymbol{\beta}}(\omega)$. One can check whether or not a particular element of $\widehat{\boldsymbol{\beta}}(\omega)$ is zero by inserting $\widehat{\boldsymbol{\beta}}(\omega)$ in for $\boldsymbol{\beta}(\omega)$, but with the particular element zeroed out and the resulting vector rescaled to be of unit length, into (9).

Significance thresholds for the smoothed spectral envelope estimate can easily be computed using the following approximations. Using the first-order Taylor expansion we have

$$\log\widehat{\lambda}(\omega) \approx \log\lambda(\omega) + \frac{\widehat{\lambda}(\omega) - \lambda(\omega)}{\lambda(\omega)},$$

[5] The Walsh functions are a completion of the Haar functions; a summary of their use in statistics is given in the works of Stoffer (1991).

so that $(n, m \to \infty)$

$$\nu_m[\log \widehat{\lambda}(\omega) - \log \lambda(\omega)] \sim \mathrm{AN}(0, 1). \tag{10}$$

It also follows that $E[\log \widehat{\lambda}(\omega)] \approx \log \lambda(\omega)$ and $\mathrm{var}[\log \widehat{\lambda}(\omega)] \approx \nu_m^{-2}$. If there is no signal present in a sequence of length n, we expect $\lambda(j/n) \approx 2/n$ for $1 < j < n/2$, and hence approximately $(1 - \alpha) \times 100\%$ of the time, $\log \widehat{\lambda}(\omega)$ will be less than $\log(2/n) + (z_\alpha/\nu_m)$ where z_α is the $(1 - \alpha)$ upper tail cutoff of the standard normal distribution. Exponentiating, the α critical value for $\widehat{\lambda}(\omega)$ becomes $(2/n)\exp(z_\alpha/\nu_m)$. Although this method is a bit crude, from our experience, thresholding at very small α-levels (say, $\alpha = 10^{-4}$–10^{-6}, depending on the size of n) works well.

2.3. Data analysis

As a simple example, consider the sequence data presented in the study by Whisenant et al. (1991), which were used in an analysis of a human Y-chromosomal DNA fragment; the fragment is a string of length $n = 4156$ bp. The sample spectral envelope of the sequence is plotted in Fig. 3, where frequency is measured in cycles per bp. The spectral envelope can be interpreted as the *largest proportion of the total variance* at frequency ω that can be obtained for any scaling of the DNA sequence. The graph can be inspected for peaks by employing the approximate null probabilities previously given. In Fig. 3, we show the approximate 0.00001 null significance threshold for a single a priori-specified frequency ω. The null significance value was chosen small in view of the problem of making simultaneous inferences about the value of the spectral envelope over more than one frequency.

Figure 3 shows a major peak near the zero frequency, indicating that the process has long memory. Long memory is typically seen in the analysis of long DNA sequences, and the implication of this was discussed by Maddox (1992). The estimated optimal

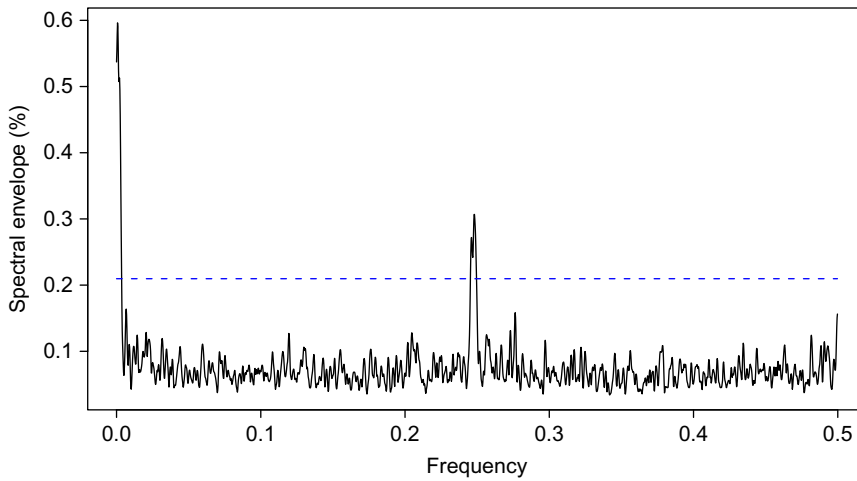

Fig. 3. Spectral envelope of a human Y-chromosomal fragment based on a smoothed periodogram estimate; the horizontal dashed line is an approximate 0.00001 significance threshold.

scalings at the zero frequency estimate are A = 0.67, C = 0.74, G = 0.03, T = 0; this particular scaling suggests that the long memory is in terms of the amino-keto alphabet (A = C, G = T). There is also a secondary peak at approximately $\omega = 0.25$ cycle per bp with a corresponding sample scaling of A = 0.41, C = 0.43, G = 0.80, T = 0. Again we see the amino pairing A = C, but in this case, G and T are different.

3. Local spectral envelope

3.1. Piecewise stationarity

Long DNA sequences are heterogeneous and hence there is a need to establish methods to investigate local behavior. In particular, as discussed in the Introduction, the genetic model is that CDS are segments of DNA that are dispersed throughout the sequence and separated by regions of noncoding or noise. Because genetic information is contained in segments, piecewise stationarity appears to be a suitable model.

A $k \times 1$ vector-valued *piecewise stationary process*, $\{\boldsymbol{Y}_{s,n}\}_{s=0}^{n-1}$, for $n \geq 1$, is defined to be

$$\boldsymbol{Y}_{s,n} = \sum_{b=1}^{B} \boldsymbol{Y}_{s,b}\, \mathcal{I}(s/n, U_b), \tag{11}$$

where the $\boldsymbol{Y}_{s,b}$ are stationary processes with continuous $k \times k$ spectral matrices $f_{Y,b}(\omega)$, and where $U_b = [u_{b-1}, u_b) \subset [0, 1)$ is an interval, and $\mathcal{I}(s/n, U_b)$ is an indicator that takes the value 1 if $s/n \in U_b$, and 0 otherwise. For ease of notation, we rescale time in each block so that

$$\{\boldsymbol{Y}_{s,b}\colon s/n \in U_b\} \mapsto \{\boldsymbol{Y}_{t,b}\colon t = 1, \ldots, n_b\},$$

where the number of observations in segment b is n_b and $\sum_{b=1}^{B} n_b = n$. This rescaling of time represents a simple time shift to the origin wherein $\boldsymbol{Y}_{s,b} \mapsto \boldsymbol{Y}_{t,b}$ for $s/n \in U_b$ with $t = s + 1 - \sum_{i=1}^{b-1} n_i$.

We shall say that a categorical time series, $\{X_{s,n}\}$, on a finite state space and with nonzero marginal probabilities (as discussed in Section 1), is *piecewise stationary* if the corresponding $k \times 1$ points process, $\{\boldsymbol{Y}_{s,n}\}$, is piecewise stationary. Quite often, infill asymptotics is used for locally stationary processes (e.g., Dahlhaus, 1997). However, a DNA sequence is truly a discrete-time process, so it would be unrealistic to consider an infill asymptotic situation wherein we assume we are able to obtain more observations in a segment as the number of observations grows. In our case, we rely on increasing domain asymptotics to approximate the behavior of the estimated spectral envelope for suitably large segments. For small segments, simple Monte Carlo simulations can be used to approximate the small sample null distribution of the spectral envelope estimator.

If $X_{s,n}$ is a piecewise stationary categorical time series, we define the local spectral as the local analog of the optimality criterion described in (5), that is,

$$\lambda_b(\omega) = \sup_{\boldsymbol{\beta}} \left\{ \frac{\boldsymbol{\beta}' f_{Y,b}^{re}(\omega) \boldsymbol{\beta}}{\boldsymbol{\beta}' V_b \boldsymbol{\beta}} \right\}, \tag{12}$$

for $b = 1, \ldots, B$, where V_b is the variance–covariance matrix of $\boldsymbol{Y}_{t,b}$, which are the indicator vectors in block b as described in the previous section. Analogous to Section 1, we define $\lambda_b(\omega)$ to be the *local spectral envelope* and the corresponding eigenvector $\boldsymbol{\beta}_b(\omega)$ to be the *local optimal scaling* of block b and frequency ω.

The sample local spectral envelope is obtained analogously to the stationary case, the *local periodogram* of the data $\{\boldsymbol{Y}_{s,n} : s/n \in U_b\}$ in block b, for $b = 1, \ldots, B$, is given by

$$\widetilde{f}_b(\omega) = \boldsymbol{d}_b(\omega)\boldsymbol{d}_b^*(\omega), \tag{13}$$

where

$$\boldsymbol{d}_b(\omega) = n_b^{-1/2} \sum_{t=1}^{n_b} \boldsymbol{Y}_{t,b} \ \exp\{-2\pi i t\omega\}$$

is the finite Fourier transform of the data $\{\boldsymbol{Y}_{t,b} : t = 1, \ldots, b\}$. A smoothed estimate of the local spectral density can be obtained as

$$\widehat{f}_b(\omega_j) = \sum_{q=-m_b}^{m_b} h_{q,b}\, \widetilde{f}_b(\omega_{j+q}), \tag{14}$$

where $\omega_j = j/n_b$ and the amount and type of smoothing, $\{h_{q,b}\}$, depends on n_b among other things. Under the assumption of piecewise stationarity, and in the case that the stationary blocks are known, the results regarding estimation follow from the previous section. In particular, the results (6)–(10) apply to the local estimation case provided that n_b is sufficiently large. As previously stated, the small n_b case can be dealt with by direct simulation.

3.2. Data analysis

As a simple example of the kind of analysis that can be accomplished, we consider the gene BNRF1 (bp 1736–5689) of the Epstein–Barr virus (EBV); note that the gene is nearly 4000-bp long. Fig. 4 shows a dynamic spectral envelope with a block size of $n_b = 500$. It is immediately evident from the figure that the even within a gene, there is heterogeneity. There is, however, a basic cyclic pattern that exists through most of the gene as evidenced by the peak at $\omega = 1/3$ except at the end of the gene. Table 1 shows the optimal scalings at the one-third frequency and we note that the corresponding alphabets are somewhat consistent in the "significant" blocks, with each block indicating a weak–strong bonding alphabet (A = T, C = G), except block number five (bp 3736-4235).

3.3. Dyadic segmentation

Next, we discuss a systematic method for obtaining a local spectral envelope. The basic idea is laid out in the works done by Stoffer et al. (2002) and is made more rigorous by Jeong (2011). In the study by Stoffer et al. (2002), we presented asymptotic theory

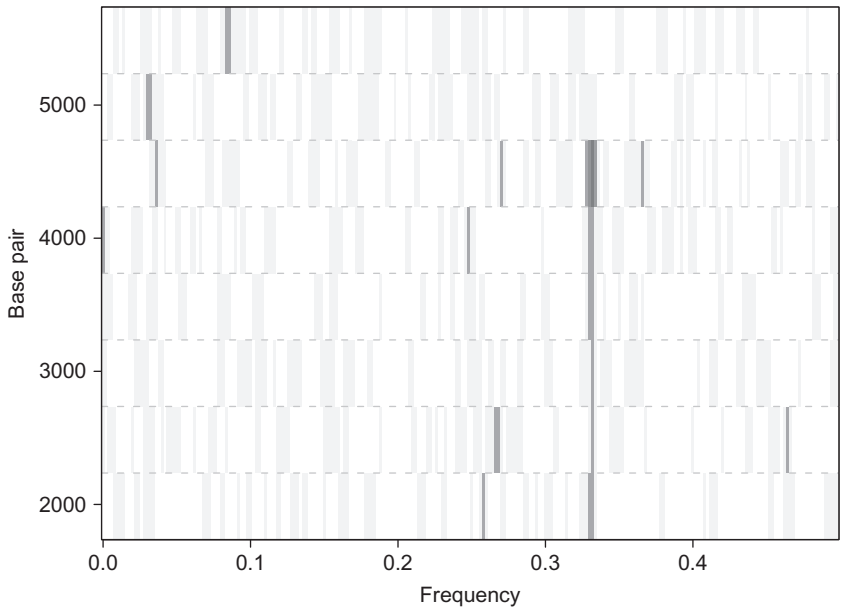

Fig. 4. Dynamic spectral envelope estimates for the BNRF1 gene (bp 1736–5689) of the Epstein–Barr virus (EBV). The horizontal dashed lines indicate the blocks, and values over the approximate 0.005 null significance threshold are indicated by darker regions.

Table 1
Blockwise optimal scaling, $\widehat{\boldsymbol{\beta}}(1/3)$, for the Epstein–Barr BNRF1 gene example

Block (bp)	A	C	G	T
1736–2235	0.26	0.69	0.68	0
2236–2735	0.23	0.71	0.67	0
2736–3235	0.16	0.56	0.82	0
3236–3735	0.15	0.61	0.78	0
3736–4235	0.30	0.35	0.89	0
4236–4735	0.22	0.61	0.76	0
4736–5235[a]	0.41	0.56	0.72	0
5236–5689[a]	0.90	−0.43	−0.07	0

[a] $\widehat{\lambda}(1/3)$ is not significant in this block.

in the local stationary case that is similar to stationary case, with the added condition that the block size is large. One problem that was not considered in that article was whether local spectral estimates were independent across blocks. To this end, we state the following lemma wherein we assume the conditions under which Stoffer et al. (2002, Theorem 3.3) hold; these conditions ensure the asymptotic normality of the local spectral envelope, $\lambda_b(\omega)$ defined in (12). The main condition for which the independence holds is if the process is not long memory; as was seen in the previous section,

this assumption is likely to be violated in long DNA sequences, but perhaps not in relatively short subsequences. In the following lemma, we define $\omega_{j:n} = j_n/n$, where $\{j_n\}$ is a sequence of integers chosen so that j_n/n is the closest Fourier frequency to the frequency of interest, ω; i.e., $\omega_{j:n} \to \omega$ as $n \to \infty$.

LEMMA 1. *Let* $\{X_t\}$ *be stationary with covariance function,* $\gamma(h)$*, satisfying* $\sum_h |h|\,|\gamma(h)| < \infty$*. Suppose we observe* $X_1, \ldots, X_n, X_{n+1}, \ldots X_{2n}$ *for* $n \geq 1$*. For* $j = 0, 1, \ldots, n-1$*, let*

$$d_1(\omega_{j:n}) = n^{-1/2} \sum_{t=1}^{n} X_t \exp(-2\pi i t \omega_{j:n})$$

and let

$$d_2(\omega_{k:n}) = n^{-1/2} \sum_{t=1}^{n} X_{t+n} \exp(-2\pi i t \omega_{k:n}),$$

such that $\omega_{j:n} \to \omega_1$ *and* $\omega_{k:n} \to \omega_2$*. Then* $d_\ell(\cdot)$*, for* $\ell = 1, 2$*, is asymptotically* $(n \to \infty)$ *complex normal and such that* $d_1(\omega_{j:n})$ *and* $d_2(\omega_{k:n})$ *are asymptotically independent for any* ω_1 *and* ω_2*.*

PROOF. The asymptotic complex normality follows directly from the works of Shumway and Stoffer (2011, Theorem C.4). The independence follows from the following inequality.

$$\begin{aligned}
|\mathrm{cov}\{d_1(\omega_{j:n}), d_2(\omega_{k:n})\}| &\leq n^{-1} \sum_{t=1}^{n} \sum_{s=1}^{n} |\gamma(n+s-t)| \\
&= \sum_{j=1}^{n} \frac{j}{n} |\gamma(j)| + \sum_{j=n+1}^{2n-1} \frac{2n-j}{n} |\gamma(j)| \\
&\leq \sum_{j=1}^{n} \frac{j}{n} |\gamma(j)| + \sum_{j=n+1}^{2n-1} |\gamma(j)|.
\end{aligned}$$

Let $n \to \infty$, then by Kronecker's Lemma, $\sum_{j=1}^{n} \frac{j}{n} |\gamma(j)| \to 0$, and by the absolute summability of $\gamma(h)$, we have $\sum_{j=n+1}^{2n-1} |\gamma(j)| \to 0$. □

Based on Lemma 1, and provided that we are not in the long memory case, comparison of blockwise spectral envelope estimates and corresponding scalings is readily available based on the asymptotic independence of the estimates in each block. For example, a large sample test of equality of $\boldsymbol{\beta}$s at the same frequency in two blocks would proceed as follows. Let $\widehat{\boldsymbol{\beta}}_1$ is estimate of the optimal scale at frequency ω_j in the first block, $\widehat{\boldsymbol{\beta}}_2$ in the second block (with the first nonzero element of β positive) and $\widehat{\Sigma}_i$

are the estimates of the corresponding covariance matrices given in (8). Under the null hypothesis that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ (which does not imply $\Sigma_1 = \Sigma_2$) we have that

$$\nu_m^2 \left(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2\right)' \left(\widehat{\Sigma}_1 + \widehat{\Sigma}_2\right)^{-1} \left(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2\right) \dot{\sim} \chi_k^2;$$

recall k is one less than the size of the alphabet.

Now, we discuss a tree-based adaptive segmentation method for finding the blocks $b = 1, \ldots, B$. The strategy is to divide the sequence into small blocks and then to recombine adjacent blocks whose estimated local spectral envelopes are sufficiently similar. The basic idea is that adjacent blocks with similar local spectral envelope estimates give similar genetic information. The main feature of the algorithm is it divides the sequence in a dyadic manner using a measure of distance (or discrepancy) between the genetic coding information contained at two adjacent blocks. The algorithm, which was inspired by Adak (1998) and was suggested by Stoffer et al. (2002), is as follows.

1. *Set the maximum level* J. The value of J determines the smallest possible size of the segmented blocks. For a sequence of length n, the smallest blocks have length $n/2^J$. Ideally, the block sizes should be small enough so that one can separate useful genetic information unique to that block from the noncoding material (noise). One should be careful, however, about making the blocks too small. Blocks have to be large enough to give good estimates of the local spectral envelope. Our recommendation is that the block size should be at least 2^8.
2. *Form the blocks.* At each level $j = 0, \ldots, J$, divide the data sequence into 2^j blocks. Denote $B(j, \ell)$ to be the ℓth block on level j, where $\ell = 1, \ldots, 2^j$. The first block on level j is denoted as $B(j, 1)$ and the last as $B(j, 2^j)$. The "inner" blocks at level j are $B(j, \ell)$, (where $\ell = 2, \ldots, 2^j - 1$). For any level $j = 0, \ldots, J$, block $B(j, \ell)$, for $\ell = 1, \ldots, 2^j$, consists of the $M_j = n/2^j$ elements $\{X_{[(\ell-1)n/2^j]}, \ldots, X_{[\ell n/2^j - 1]}\}$.
3. *Estimate the spectral envelope.* Compute an estimate of the local spectral envelope, $\widehat{\lambda}_{j,\ell}(\omega_k)$, at each fundamental frequency $\omega_k = k/M_j$ ($k = 0, \ldots, M_j/2$) in each block $B(j, \ell)$ where $j = 0, \ldots, J$, and $\ell = 1, \ldots, 2^j$.
4. *Create a table of distances.* Let $\delta[\cdot, \cdot]$ be a distance (discrepancy) measure between the spectral envelope estimates of two children blocks. We will discuss choosing such a measure after the algorithm is presented. Using the distance measure, create a table of distances corresponding to each block, $B(j, \ell)$, namely,

   $$D(j, \ell) = \delta[\widehat{\lambda}_{j+1,2\ell-1}(\omega), \widehat{\lambda}_{j+1,2\ell}(\omega)],$$

   for $\ell = 1, \ldots, 2^j$, and for each level $j < J$.
5. *Mark the blocks for final segmentation.* Mark all the blocks $B(J-1, \ell)$, at level $J-1$ for $\ell = 1, \ldots, 2^{J-1}$. For $j = J-2$, and $\ell = 1, \ldots, 2^j$, if

   $$D(j, \ell) \le D(j+1, 2\ell - 1) + D(j+1, 2\ell),$$

   then mark the block $B(j, \ell)$ and leave $D(j, \ell)$ unchanged. Otherwise, leave the block $B(j, \ell)$ as unmarked and set

   $$D(j, \ell) = D(j+1, 2\ell - 1) + D(j+1, 2\ell).$$

Iterate this procedure for $j = J-3, J-4, \ldots, 0$. The *final segmentation* of the DNA sequence is the set of highest marked blocks $B(j,\ell)$ such that $B(j,\ell)$ is marked and its parent block and ancestor blocks are not marked.

6. *Classification.* For the final segmentation, use the information in the estimated local spectral envelope to classify a segment as (i) highly likely to contain CDS, (ii) highly likely to contain noncoding, or (iii) uncertain. A specific classification method is discussed below.

As opposed the recommendation in the works of Stoffer et al. (2002), our preferred choice for a distance measure in Step 4 is a symmetric Kullback–Leibler divergence between the local spectral envelope in children blocks $B(j+1; 2\ell-1)$ and $B(j+1; 2\ell)$. To this end, we define the distance measure,

$$D(j,\ell) = \frac{1}{M_j/2} \sum_{j=1}^{M_j/2} \left[\widehat{\lambda}_{j+1,2\ell-1}(\omega_j) - \widehat{\lambda}_{j+1,2\ell}(\omega_j)\right] \log \frac{\widehat{\lambda}_{j+1,2\ell-1}(\omega_j)}{\widehat{\lambda}_{j+1,2\ell}(\omega_j)}, \qquad (15)$$

where $\omega_j = j/M_j$. The use of the measure is discussed in the study by Jeong (2011), where it is shown that, if $\lambda_{j+1,2\ell-1}(\omega) = \lambda_{j+1,2\ell}(\omega)$, then, as $M_j \to \infty$,

$$\Pr\{D(j,\ell) > D(j+1,2\ell-1) + D(j+1,2\ell)\} \to 0.$$

In other words, for large block sizes, the probability that the algorithm splits a block when in fact the block should not be split is small. Once the final segmentation is determined, a *classification rule* should be put into place on the segmented sequence. Such rules are perhaps best left to molecular biologists and should take into account the type of DNA being considered. Experience with viruses leads us to the following classification rule for viruses, which we demonstrate in an example. To this end, (i) a block is designated as containing only coding if the local estimated spectral envelope exhibits a peak at frequency 1/3 and other nonzero frequencies such as the 1/10 frequency predicted by Trifonov in the early 1980s. (ii) A block is designated as containing both coding and noncoding if the spectral envelope exhibits a peak at (or near) the zero frequency as well as a peak at frequency 1/3, and possibly other nonzero frequencies. (iii) A block is designated as containing noncoding (noise) if the spectral envelope is either flat, indicating white noise, or has a peak at, or near, the zero frequency and no other peaks, indicating fractional noise. (iv) A block is designated as containing other interesting features (e.g., repeat regions) if spectral envelope exhibits several nonzero peaks. (v) If adjacent blocks are classified in the same way, and the optimal scaling indicates the same alphabets, they may be recombined.

3.4. Data analysis

As an example, we present the analysis of a subsequence of the Epstein–Barr virus genome. The subsequence consists of bp 46001–54192; the length of the series is $n = 2^{13} = 8192$. Table 2 shows a portion of the EMBL file on the virus, and there are three interesting regions indicated within this subsequence. The segment contains two coding sequences (CDS), one from bp 46,333 to 47481 [BWRF1], and another from

Table 2
Section of the Epstein–Barr file at the European Molecular Biology Laboratory (EMBL)

Key	Location/Qualifiers	Key	Location/Qualifiers
CDS[a]	46333..47481 /note="BWRF1 reading frame 12"	mRNA	49852..50032 /note="exon (Bodescot et al., 1984)
misc_feature	47007..47007 /note="BAM: BamH1 W/Y"	misc_feature	50003..50003 /note="polyA signal: AATAAA, end of
mRNA	47761..47793 /note="Exon Y1 Bodescot et al., 1984"		T1 RNA and EBNA-2 RNA (3.0kb latent RNA in IB4 cells)"
promoter	47831..47831 /note="TATA: TATAAGT"	promoter	complement(50156..50156) /note="TATA: TATAAGT"
mRNA	47878..47999 /note="Exon Y2 Bodescot et al., 1984 EBNA-1 (Speck and Strominger, 1985) last common exon"	misc_feature	complement(50317..50317) /note="polyA signal: AATAAA, early RNA from 52817"
		repeat_region[a]	50578..52115
misc_feature	complement(48023..48023) /note="polyA signal: AATAAA"		/note="12 x "125bp" repeats"
		misc_feature	complement (50578..52557)
CDS[a]	48386..50032		/note="BHLF1 early reading frame"
	/note="Coding exon for EBNA-2 (Sample et al., 1986)"	misc_feature	52654..53697 /note="region homologous to Eco
mRNA	48386..48444 /note="exon Bodescot et al., 1984"	promoter	complement(52817..52817) /note="TATA: GATAAAA early RNA containing
CDS[a]	48429..49964 /note="BYRF1, encodes EBNA-2 (Dambaugh et al., 1984; Dillner et al., 1984)"	promoter	BHLF1 (Jeang and Hayward, 1983; Freese et al., 1983)" 53759..53759

[a] Indicates interesting regions of the sequence.

bp 48386 to 50032 [BYRF1]. Also notable is a large repeat region from bp 50578 to 52115; repeat regions are highly repetitive regions DNA. Repeat regions are as much of an interest to molecular biologists as CDS. For example, in humans, repeat regions are often associated with disease syndromes. In this example, we set the lowest level at $J = 5$ so that the smallest blocks have 256 elements.

The table of distances indicating the best segmentation and classifications is shown in Table 3. We note that adjacent blocks with the same classification can be recombined; this situation happens with blocks $B(4, 10)$ and $B(3, 6)$. The spectral envelopes for the final segmentation are displayed in Fig. 5. The algorithm locates the interesting regions of the DNA sequence considered here. In particular, block $B(1, 1)$, indicates a CDS in the subsequence. The spectral envelope of that region has peaks at the predicted frequencies of $1/10$ and $1/3$. The combination of blocks $B(4, 10)$ and $B(3, 6)$,

Table 3
Distances (in %) as defined in (15) for each block

Level	$B(j,\ell)$															
$j=0$	$28 \mapsto 25$															
$j=1$	9 **[C]**								$29 \mapsto 16$							
$j=2$	7				7				$30 \mapsto 8$				8 **[N]**			
$j=3$	6		7		15		11		$19 \mapsto 6$		2 **[R]**		7		8	
$j=4$	6	3	7	7	6	14	19	5	6 **[N]**	0 **[R]**	3	3	6	46	12	6

Note: The $\mapsto$ symbol indicates that an original distance has been reset, and thus a block without a $\mapsto$ symbol indicates a marked block, all according to step 5 of the algorithm. The best segmentation is marked with a letter indicating the classification: [**C**] = CDS, [**N**] = Noise, [**R**] = Repeat region.

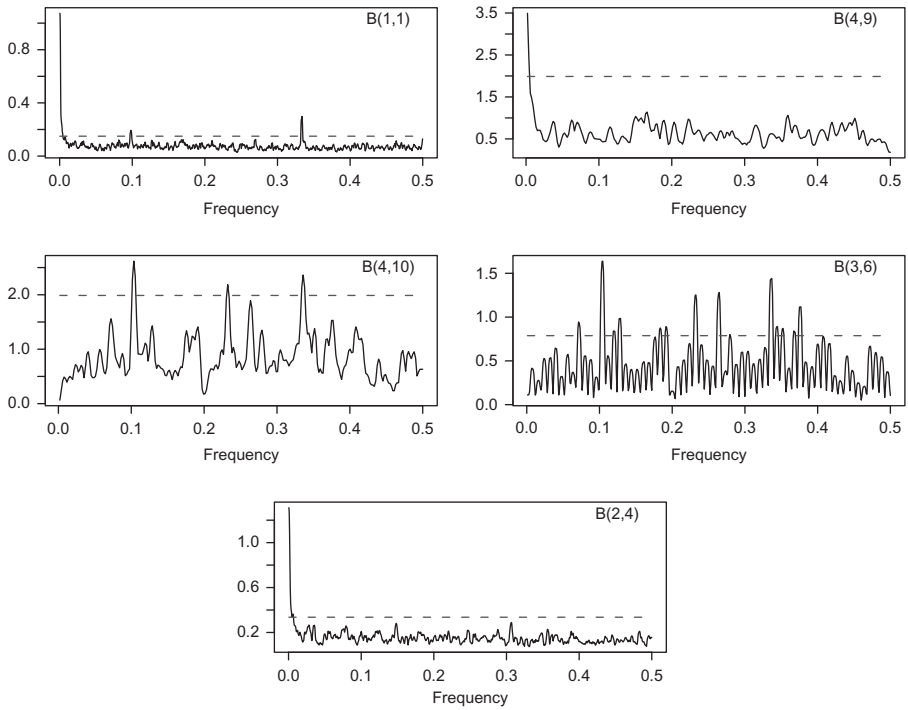

Fig. 5. Spectral envelopes for the DNA subsequence in each block of best segmentation and an approximate 0.00001 significance threshold shown as a dashed line.

which includes bp 50609–52144, correctly identifies a large repeat region (the actual location is bp 50578–52115); notice the difference between a CDS region and a repeat region, which has multiple peaks. It is of course reasonable to recombine these two blocks. Finally, the spectral envelopes in blocks $B(4,9)$ and $B(2,4)$ are similar and

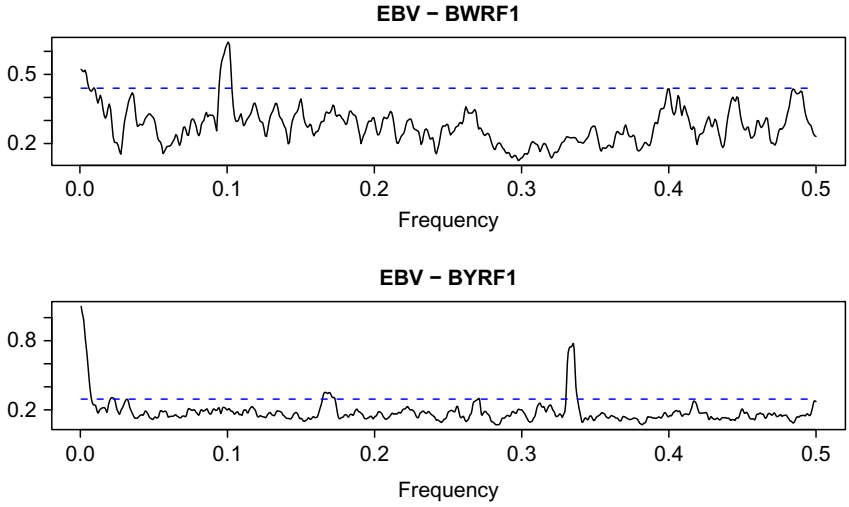

Fig. 6. Spectral envelopes of the BWRF1 gene and the BYRF1 gene that are both within block $B(1,1)$. An approximate 0.00001 significance threshold is shown as a dashed line.

both indicate fractional noise. Figure 5 includes the approximate 0.00001 null significance thresholds, shown as dashed lines, for reference. Note that the coding sequences, which are both within block $B(1,1)$, are not separated by the algorithm. This is likely due to the fact that the CDS BWRF1 is in both blocks $B(2,1)$ and $B(2,2)$, which were combined in the algorithm. Although the genes are not separated by the algorithm, it correctly identifies a region of interest. Further investigation of the region of interest indicates that there are two different mechanisms, one indicated by the $1/10$ frequency and the other by the $1/3$ frequency. The spectral envelopes of the individual genes are displayed in Fig. 6, where the distinction between the two genes are evident.

3.5. Discussion

Before closing this section, I will mention that smoothing the spectral matrix of a multivariate process takes special care. For example, all the elements of the periodogram matrix are smoothed the same way with the typical smoothing technique given in (3). This is done to ensure that the estimate, $\widehat{f}(\omega)$, is non-negative definite. There are, however, many situations, including the analysis DNA sequences, for which different components of the spectral matrix have different degrees of smoothness. To overcome this problem, Rosen and Stoffer (2007) proposed a Bayesian approach that used Markov chain Monte Carlo techniques to fit smoothing splines to each component, real and imaginary, of the Cholesky decomposition of the periodogram matrix. The spectral estimator is then obtained by reconstructing the spectral estimator from the smoothed Cholesky decomposition components. The technique produces an automatically smoothed spectral matrix estimator along with samples from the posterior distributions of the parameters to facilitate inference. We will not present an analysis using the technique here, but interested readers can see the DNA sequence example in Section 3.3 of the paper by Rosen and Stoffer (2007).

4. Detection of genomic differences

As discussed in the Introduction, the problem of matching two DNA sequences is of essential interest to molecular biologists. The paper of Waterman and Vingron (1994), which gives some background to problem, is written for statisticians. They noted that new DNA or protein sequences are compared with one or more sequence databases to find similar or homologous sequences that have already been studied. Moreover, there are numerous examples of important discoveries resulting from these comparisons. For example, when the cystic fibrosis gene was cloned and sequenced, a database search revealed that the gene product had similarity to a family of related ATP-binding proteins involved in active transport of small hydrophilic molecules across the cytoplasmic membrane (Riordan et al., 1989).

4.1. The general problem

Stoffer and Tyler (1998) discussed a more general problem and we give some background here. In the general case, X_{1t} and X_{2t}, $t = 0, \pm 1, \pm 2, \ldots$, are categorical sequences taking values in possibly different state spaces of dimensions $k_1 + 1$ and $k_2 + 1$, respectively. Consider two nonconstant transformations g and h with $g(X_{1t})$ and $h(X_{2t})$ being real-valued time series such that $g(X_{1t})$ has continuous spectral density $f_{gg}(\omega)$ and $h(X_{2t})$ has continuous spectral density $f_{hh}(\omega)$. We denote the complex-valued cross-spectral density of the two series $g(X_{1t})$ and $h(X_{2t})$ by $f_{gh}(\omega)$. A measure of the degree of similarity between the sequences $g(X_{1t})$ and $h(X_{2t})$ at frequency ω is the squared coherency,

$$\rho^2_{gh}(\omega) = \frac{|f_{gh}(\omega)|^2}{f_{gg}(\omega) f_{hh}(\omega)}. \tag{16}$$

Of course the value of $\rho^2_{gh}(\omega)$ will depend on the choices of the transformations g and h. If X_{1t} and X_{2t} are independent, then so are $g(X_{1t})$ and $h(X_{2t})$, for any g and h, in which case $\rho^2_{gh}(\omega) = 0$ for all ω. The main goal here is to find g and h, under various constraints, to maximize the squared coherency $\rho^2_{gh}(\omega)$. If the maximized value of $\rho^2_{gh}(\omega)$ is small, we can say that the two sequences X_{1t} and X_{2t} do not match at frequency ω. If the maximized value of $\rho^2_{gh}(\omega)$ is large, then the resulting transformations g and h can help in understanding the nature of the similarity between the two sequences.

To this end, identify the categorical sequence X_{1t} with the vector indicator process $\boldsymbol{Y}_{1t}$, where $\boldsymbol{Y}_{1t}$ is a $k_1 \times 1$ vector with a one in the jth position if X_{1t} is in state j ($j = 1, \ldots, k_1$) at time t and zeros elsewhere. If X_{1t} is in state $k_1 + 1$, then $\boldsymbol{Y}_{1t}$ is the zero vector. Similarly, we identify X_{2t} with the $k_2 \times 1$ vector indicator process $\boldsymbol{Y}_{2t}$. We assume the existence of the $k_i \times k_i$ ($i = 1, 2$), nonsingular spectral matrices $f_{11}(\omega)$ and $f_{22}(\omega)$ of $\boldsymbol{Y}_{1t}$ and $\boldsymbol{Y}_{2t}$, respectively, and denote the $k_1 \times k_2$ cross-spectral matrix between $\boldsymbol{Y}_{1t}$ and $\boldsymbol{Y}_{2t}$ by $f_{12}(\omega)$.

To describe the problem in terms of scaling sequences, let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{k_1})' \in \mathbb{R}^{k_1}$, $\boldsymbol{\alpha} \neq \mathbf{0}$, be a vector of reals (scalings) associated with the categories of the first sequence, X_{1t}, and let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{k_2})' \in \mathbb{R}^{k_2}$, $\boldsymbol{\beta} \neq \mathbf{0}$, be a vector of scalings associated with the

categories of the second sequence, X_{2t}. Define the real-valued series

$$\begin{aligned} X_{1t}(\boldsymbol{\alpha}) &= \alpha_j \quad \text{if } X_{1t} \text{ is in state } j \text{ for } j = 1, \ldots, k_1, \\ X_{2t}(\boldsymbol{\beta}) &= \beta_j \quad \text{if } X_{2t} \text{ is in state } j \text{ for } j = 1, \ldots, k_2, \end{aligned} \tag{17}$$

where, in addition, $X_{1t}(\boldsymbol{\alpha}) = 0$ if X_{1t} is in state $k_1 + 1$, and $X_{2t}(\boldsymbol{\beta}) = 0$ if X_{2t} is in state $k_2 + 1$. Since the scaled series can be written as $X_{1t}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}' \boldsymbol{Y}_{1t}$, and $X_{2t}(\boldsymbol{\beta}) = \boldsymbol{\beta}' \boldsymbol{Y}_{2t}$, the squared coherency between $X_{1t}(\boldsymbol{\alpha})$ and $X_{2t}(\boldsymbol{\beta})$ can be written as

$$\rho^2_{12}(\omega; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{|\boldsymbol{\alpha}' f_{12}(\omega) \boldsymbol{\beta}|^2}{[\boldsymbol{\alpha}' f^{re}_{11}(\omega) \boldsymbol{\alpha}]\, [\boldsymbol{\beta}' f^{re}_{22}(\omega) \boldsymbol{\beta}]}. \tag{18}$$

Setting $\boldsymbol{a} = f^{re}_{11}(\omega)^{1/2} \boldsymbol{\alpha}$ and $\boldsymbol{b} = f^{re}_{22}(\omega)^{1/2} \boldsymbol{\beta}$, subject to $\boldsymbol{a}'\boldsymbol{a} = 1$ and $\boldsymbol{b}'\boldsymbol{b} = 1$, define

$$Q(\omega) = f^{re}_{11}(\omega)^{-1/2} f_{12}(\omega) f^{re}_{22}(\omega)^{-1/2} = Q^{re}(\omega) + i\ Q^{im}(\omega) \tag{19}$$

and write (18) as

$$\rho^2_{12}(\omega; \boldsymbol{a}, \boldsymbol{b}) = [\boldsymbol{a}' Q^{re}(\omega) \boldsymbol{b}]^2 + [\boldsymbol{a}' Q^{im}(\omega) \boldsymbol{b}]^2. \tag{20}$$

The goal is to find $\boldsymbol{a}$ and $\boldsymbol{b}$ to maximize (20) for each ω of interest. Several approaches to the maximization are available; one approach is based on the following observations.

PROPOSITION 1. (Stoffer & Tyler) *Fix ω and drop it from the notation. Then (20) can be written as*

$$\rho^2_{12}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}'(Q^{re} \boldsymbol{b}\boldsymbol{b}' Q^{re} + Q^{im} \boldsymbol{b}\boldsymbol{b}' Q^{im}) \boldsymbol{a} = \boldsymbol{b}'(Q^{re} \boldsymbol{a}\boldsymbol{a}' Q^{re} + Q^{im} \boldsymbol{a}\boldsymbol{a}' Q^{im}) \boldsymbol{b}. \tag{21}$$

Let $\boldsymbol{b}_0$ be an arbitrary real-valued $k_2 \times 1$ unit length vector. Define the sequence of vectors $\boldsymbol{a}_j$ to be the eigenvector corresponding to the largest root of the at most rank 2, nonnegative definite matrix

$$Q^{re} \boldsymbol{b}_{j-1} \boldsymbol{b}'_{j-1} Q^{re'} + Q^{im} \boldsymbol{b}_{j-1} \boldsymbol{b}'_{j-1} Q^{im'} \tag{22}$$

and the sequence $\boldsymbol{b}_j$ to be the eigenvector corresponding to the largest root of the at most rank 2, nonnegative definite matrix

$$Q^{re'} \boldsymbol{a}_j \boldsymbol{a}'_j Q^{re} + Q^{im'} \boldsymbol{a}_j \boldsymbol{a}'_j Q^{im}, \tag{23}$$

for $j = 1, 2, \ldots$ Then, from the first part of (21) it follows that $\rho^2(\boldsymbol{a}_{j+1}, \boldsymbol{b}_j) \geq \rho^2(\boldsymbol{a}, \boldsymbol{b}_j)$ for any $\boldsymbol{a}$ of unit length, and from the second part of (21) it follows that $\rho^2(\boldsymbol{a}_{j+1}, \boldsymbol{b}_{j+1}) \geq \rho^2(\boldsymbol{a}_{j+1}, \boldsymbol{b})$ for any $\boldsymbol{b}$ of unit length. Thus,

$$\rho^2(\boldsymbol{a}_{j+1}, \boldsymbol{b}_{j+1}) \geq \rho^2(\boldsymbol{a}_{j+1}, \boldsymbol{b}_j) \geq \rho^2(\boldsymbol{a}_j, \boldsymbol{b}_j). \tag{24}$$

The algorithm can be used to find the optimal scalings at each frequency, ω, of interest. With $\mathcal{L}[A]$ denoting the eigenvector corresponding to the largest eigenvalue of matrix A, the algorithm can be initialized by setting $\boldsymbol{b}_0$ equal to either $\mathcal{L}[Q^{re}(\omega)'Q^{re}(\omega)]$ or $\mathcal{L}[Q^{im}(\omega)'Q^{im}(\omega)]$, depending on which vector produces the larger value of (20) for arbitrary $\boldsymbol{a}$. In turn, $\boldsymbol{\alpha}(\omega)$ and $\boldsymbol{\beta}(\omega)$ can be taken proportional to $f_{11}^{re}(\omega)^{-1/2}\boldsymbol{a}(\omega)$ and $f_{22}^{re}(\omega)^{-1/2}\boldsymbol{b}(\omega)$, respectively, where $\boldsymbol{a}(\omega)$ and $\boldsymbol{b}(\omega)$ maximize (20). Note that the algorithm requires only the computation of latent roots and vectors of at most rank 2, nonnegative definite matrices, regardless of the dimension of the state-spaces. Moreover, by (24), the objective function increases with each step.

In the specific problem of comparing DNA sequences, we will be interested in using the same scaling for both sequences. The next proposition establishes a condition under which that approach is optimal.

PROPOSITION 2. (Stoffer & Tyler) *Under the notation and conditions of Proposition 1, if $k_1 = k_2 = k$ and the matrices Q^{re} and Q^{im} are symmetric, the maximum value of $\rho_{12}^2(\boldsymbol{a}, \boldsymbol{b})$ is attained when $\boldsymbol{a} = \boldsymbol{b}$.*

In the case of symmetry, the algorithm in Proposition 1 is simplified by setting $\boldsymbol{b}_0$ equal to either $\mathcal{L}[Q^{re}(\omega)^2]$ or $\mathcal{L}[Q^{im}(\omega)^2]$, depending on which vector produces the larger value of $\rho_{12}^2(\omega, \boldsymbol{b}_0)$. The sequence

$$\boldsymbol{b}_j = \mathcal{L}[Q^{re}(\omega)\boldsymbol{b}_{j-1}\boldsymbol{b}_{j-1}'Q^{re}(\omega) + Q^{im}(\omega)\boldsymbol{b}_{j-1}\boldsymbol{b}_{j-1}'Q^{im}(\omega)], \tag{25}$$

for $j = 1, 2, \ldots$, replaces the alternating sequences (22)–(23); note that $\rho_{12}^2(\omega; \boldsymbol{b}_j) \geq \rho_{12}^2(\omega; \boldsymbol{b}_{j-1})$.

The problems mentioned above are related to the canonical analysis of time series as developed by Brillinger (2001, Chapter 10). The details are described in the Appendix, but briefly, if in (17), we allow $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to be complex-valued, then the solution is that $X_{1t}(\boldsymbol{\alpha})$ and $X_{2t}(\boldsymbol{\beta})$ are the canonical variate series with maximal squared coherency being the largest eigenvalue of $f_{22}(\omega)^{-1/2} f_{21}(\omega) f_{11}(\omega)^{-1} f_{12}(\omega) f_{22}(\omega)^{-1/2}$. Although this method can be used to obtain an upper bound for the real-valued cases, it is perhaps too brutal to be used in applications to DNA sequences; one obvious problem being that it leads to the use of complex-valued scales that are different for each sequence.

4.2. Models for sequence matching

In the case of matching DNA sequences, we are interested in sequences X_{1t} and X_{2t} that are defined on the same state space, $\mathcal{S} = \{c_1, \ldots, c_{k+1}\}$. In this case, it is appropriate to choose common scalings. We consider two cases, *local alignment* where the two sequences may be in phase, and *global alignment* where the sequences may be out of phase. Henceforth, the $k \times 1$ indicator sequence corresponding to the DNA sequence, X_{it} is denoted by $\boldsymbol{Y}_{it}$, for $i = 1, 2$.

In Stoffer (1987, Section 3), I developed a number of signal-plus-noise models for discrete-valued time series. In the context of matching sequences, we may use those concepts as follows. The first model, which I will call the *local alignment* model, is

$$\boldsymbol{Y}_{it} = \boldsymbol{p}_i + \boldsymbol{S}_t + \boldsymbol{e}_{it} \tag{26}$$

where $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{ik})'$ is the vector of positive probabilities $p_{ij} = \Pr(X_{it} = c_j)$, for $i = 1, 2$ and $j = 1, \ldots, k$. In addition, $\boldsymbol{S}_t$ is *possibly* a common $k \times 1$ vector-valued series that is uncorrelated with the $k \times 1$ series $\boldsymbol{e}_{it}$, $i = 1, 2$. There may be some dependence structure between $\boldsymbol{S}_t$ and $\boldsymbol{e}_{it}$ and they may take values on different supports. If we are examining a relatively short sequence, then we may assume that $\boldsymbol{S}_t$ has $k \times k$ spectral density matrix $\boldsymbol{f}_{ss}(\omega)$, and $\boldsymbol{e}_{it}$, $i = 1, 2$, have common $k \times k$ spectra denoted by $\boldsymbol{f}_{ee}(\omega)$. It will become apparent that the method is fairly robust against the assumption of common spectra for the $\boldsymbol{e}_{it}$.

Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)' \in \mathbb{R}^k$, $\boldsymbol{\beta} \neq \boldsymbol{0}$, be a vector of scalings associated with the categories $\{c_1, \ldots, c_k\}$. As before, define the real-valued series $X_{it}(\boldsymbol{\beta}) = \beta_j$ if $X_{it} = c_j$, $j = 1, \ldots, k$, and $X_{it}(\beta) = 0$ if $X_{it} = c_{k+1}$, for $i = 1, 2$. Note that $X_{it}(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{Y}_{it} = \boldsymbol{\beta}'\boldsymbol{p}_i + \boldsymbol{\beta}'\boldsymbol{S}_t + \boldsymbol{\beta}'\boldsymbol{e}_{it}$, for $i = 1, 2$. Let $f_{11}(\omega; \boldsymbol{\beta})$ be the spectrum of scaled process $X_{1t}(\boldsymbol{\beta})$; similarly, let $f_{22}(\omega; \boldsymbol{\beta})$ denote the spectrum of $X_{2t}(\boldsymbol{\beta})$ and let $f_{12}(\omega; \boldsymbol{\beta})$ denote the cross-spectrum between $X_{1t}(\boldsymbol{\beta})$ and $X_{2t}(\boldsymbol{\beta})$. The following conditions hold:

$$\begin{aligned} f_{ii}(\omega; \boldsymbol{\beta}) &= \boldsymbol{\beta}'\{f^{re}_{ss}(\omega) + f^{re}_{ee}(\omega)\}\boldsymbol{\beta}, \quad i = 1, 2, \\ f_{12}(\omega; \boldsymbol{\beta}) &= \boldsymbol{\beta}' f^{re}_{ss}(\omega)\boldsymbol{\beta}. \end{aligned} \tag{27}$$

The coherence between $X_{1t}(\boldsymbol{\beta})$ and $X_{2t}(\boldsymbol{\beta})$ is seen to be

$$\rho_{12}(\omega; \boldsymbol{\beta}) = \frac{\boldsymbol{\beta}' f^{re}_{ss}(\omega)\boldsymbol{\beta}}{\boldsymbol{\beta}'[f^{re}_{ss}(\omega) + f^{re}_{ee}(\omega)]\boldsymbol{\beta}}. \tag{28}$$

Note that the conditions of Proposition 2 are satisfied, so that choosing common scalings is optimal here.

If there is no common signal, i.e., $f_{ss}(\omega) = 0$, then $\rho_{12}(\omega; \boldsymbol{\beta}) = 0$ for any scaling $\boldsymbol{\beta}$. Thus, the detection of a common signal can be achieved by considering the maximal coherence under the model conditions. Setting $\boldsymbol{b} = [f^{re}_{ss}(\omega) + f^{re}_{ee}(\omega)]^{1/2}\boldsymbol{\beta}$, subject to $\boldsymbol{b}'\boldsymbol{b} = 1$, write (28) as

$$\rho_{12}(\omega; \boldsymbol{b}) = \boldsymbol{b}'[f^{re}_{ss}(\omega) + f^{re}_{ee}(\omega)]^{-1/2}\, f^{re}_{ss}(\omega)\, [f^{re}_{ss}(\omega) + f^{re}_{ee}(\omega)]^{-1/2}\boldsymbol{b}. \tag{29}$$

This is an eigenvalue problem, and the maximum value of (29) is the largest scalar $\lambda(\omega)$ such that

$$[f^{re}_{ss}(\omega) + f^{re}_{ee}(\omega)]^{-1/2}\, f^{re}_{ss}(\omega)\, [f^{re}_{ss}(\omega) + f^{re}_{ee}(\omega)]^{-1/2}\boldsymbol{b}(\omega) = \lambda(\omega)\boldsymbol{b}(\omega). \tag{30}$$

The optimal scaling, $\boldsymbol{\beta}(\omega)$, is taken proportional to $[f^{re}_{ss}(\omega) + f^{re}_{ee}(\omega)]^{-1/2}\boldsymbol{b}(\omega)$. This value will maximize the coherence at frequency ω between the two sequences, with the maximum value being $\lambda(\omega)$. That is, $\rho_{12}(\omega; \boldsymbol{\beta}) \leq \rho_{12}(\omega; \boldsymbol{\beta}(\omega)) = \lambda(\omega)$, with equality only when $\boldsymbol{\beta}$ is proportional to $\boldsymbol{\beta}(\omega)$. Estimation proceeds in an obvious way: Given consistent estimates $\widehat{f}_{ij}(\omega)$, for $i, j = 1, 2$, put

$$\widehat{f}^{re}_{ss}(\omega) = [\widehat{f}^{re}_{12}(\omega) + \widehat{f}^{re}_{21}(\omega)]/2 \text{ and } \widehat{f}^{re}_{ss}(\omega) + \widehat{f}^{re}_{ee}(\omega) = [\widehat{f}^{re}_{11}(\omega) + \widehat{f}^{re}_{22}(\omega)]/2. \tag{31}$$

A frequency-based test for a common signal in the scaled sequences $X_{1t}(\boldsymbol{\beta})$ and $X_{2t}(\boldsymbol{\beta})$ was described by Stoffer and Tyler (1998), the null hypothesis being that

$f_{ss}(\omega) = 0$. The basic requirement is that we smooth the periodograms by simple averaging; that is, the weights in (3) are all equal to $1/L$, where $L = 2m + 1$. In this case, it was shown that the estimated coherence based on (31) is (we use a bar over the estimates to indicate simple averaging)

$$\overline{\rho}_{12}(\omega_j;\boldsymbol{\beta}) = \frac{\boldsymbol{\beta}'\overline{f}^{re}_{ss}(\omega_j)\boldsymbol{\beta}}{\boldsymbol{\beta}'\overline{f}^{re}_{ss}(\omega_j)\boldsymbol{\beta} + \boldsymbol{\beta}'\overline{f}^{re}_{ee}(\omega_j)\boldsymbol{\beta}} = \frac{F(\omega_j;\boldsymbol{\beta}) - 1}{F(\omega_j;\boldsymbol{\beta}) + 1}, \tag{32}$$

provided $\overline{\rho}_{12}(\omega_j;\boldsymbol{\beta}) \neq 1$. Here, ω_j is a fundamental frequency, and for a fixed value of ω_j and $\boldsymbol{\beta}$, $F(\omega_j;\boldsymbol{\beta})$ has an asymptotic ($n \to \infty$) F-distribution with $2L$ numerator and denominator degrees of freedom. It follows that the scaling, say $\overline{\boldsymbol{\beta}}(\omega_j)$, that maximizes (32) also maximizes $F(\omega_j;\boldsymbol{\beta})$. Moreover, the maximum value of $F(\omega_j;\boldsymbol{\beta})$ under model (26) is $\lambda_F(\omega_j) = [1 + \overline{\lambda}(\omega_j)]/[1 - \overline{\lambda}(\omega_j)]$, where $\overline{\lambda}(\omega_j)$ denotes the sample spectral envelope for this model with estimates based on simple averaging. Note that $\lambda_F(\omega_j) = \sup F(\omega_j;\boldsymbol{\beta})$, over $\boldsymbol{\beta} \neq \mathbf{0}$. Under the assumption that $\boldsymbol{Y}_{1t}$ and $\boldsymbol{Y}_{2t}$ are mixing, the asymptotic ($n \to \infty$) null distribution of $\lambda_F(\omega_j)$ is that of Roy's largest root. Finite sample null distributions under the additional model assumption that $\boldsymbol{e}_{1t}$ and $\boldsymbol{e}_{2t}$ are both white noise can be obtained by direct simulation. Details can be found in the works of Stoffer and Tyler (1998).

The model can be extended to include the possibility that there may be many signals common to each sequence, and that the sequences are not necessarily aligned. The general *global alignment* model is

$$\boldsymbol{Y}_{1t} = \boldsymbol{p}_1 + \sum_{j=1}^{q} \boldsymbol{S}_{jt} + \boldsymbol{e}_{1t} \quad \text{and} \quad \boldsymbol{Y}_{2t} = \boldsymbol{p}_2 + \sum_{j=1}^{q} \boldsymbol{S}_{j,t-\tau_j} + \boldsymbol{e}_{2t}, \tag{33}$$

where $\boldsymbol{S}_{jt}$, $j = 1, \ldots, q$, are zero-mean realizations of stationary $k \times 1$ vector-valued time series that are mutually uncorrelated, and in addition are uncorrelated with the zero-mean, stationary $k \times 1$ vector-valued series $\boldsymbol{e}_{1t}$ and $\boldsymbol{e}_{2t}$. Furthermore, $\boldsymbol{S}_{jt}$ has $k \times k$ spectral density matrix $f_{S_j}(\omega)$, $j = 1, \ldots, q$, and $\boldsymbol{e}_{it}$, $i = 1, 2$, have common $k \times k$ spectra denoted by $f_{ee}(\omega)$. Again, it will become apparent that the estimation procedure is robust against the assumption of equal spectra.

There is no need to specify the phase shifts, $\tau_1, \ldots, \tau_q$, or the integer $q \geq 0$, however, the problem of their estimation is interesting. We consider the following method to help decide whether or not $q = 0$. First, note that if $q > 0$, then

$$f_{11}(\omega) = f_{22}(\omega) = \sum_{j=1}^{q} f_{S_j}(\omega) + f_{ee}(\omega), \quad \text{and} \quad f_{12}(\omega) = \sum_{j=1}^{q} f_{S_j}(\omega)\exp(i\omega\tau_j). \tag{34}$$

Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)' \in \mathbb{R}^k$, $\boldsymbol{\beta} \neq 0$, be a vector of scalings, write $X_{it}(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{Y}_{it}$, for $i = 1, 2$, so that the squared coherency between $X_{1t}(\boldsymbol{\beta})$ and $X_{2t}(\boldsymbol{\beta})$ is

$$\rho^2_{12}(\omega;\boldsymbol{\beta}) = \frac{\left|\sum_{j=1}^{q} \boldsymbol{\beta}' f^{re}_{S_j}(\omega)\boldsymbol{\beta}\ \exp(i\omega\tau_j)\right|^2}{\left|\boldsymbol{\beta}' f^{re}(\omega)\boldsymbol{\beta}\right|^2}, \tag{35}$$

where $f(\omega) = f_{11}(\omega) = f_{22}(\omega)$. Setting $\boldsymbol{b} = f^{re}(\omega)^{1/2}\boldsymbol{\beta}$, with the constraint $\boldsymbol{b}'\boldsymbol{b} = 1$, write (35) as

$$\rho_{12}^2(\omega;\boldsymbol{b}) = \left|\boldsymbol{b}'\left\{\sum_{j=1}^{q} f^{re}(\omega)^{-1/2} f_{S_j}^{re}(\omega) f^{re}(\omega)^{-1/2}\exp(i\omega\tau_j)\right\}\boldsymbol{b}\right|^2. \tag{36}$$

Define the complex-valued matrix $Q(\omega)$ as

$$Q(\omega) = \sum_{j=1}^{q} f^{re}(\omega)^{-1/2} f_{S_j}^{re}(\omega) f^{re}(\omega)^{-1/2}\exp(i\omega\tau_j) = Q^{re}(\omega) + i\ Q^{im}(\omega), \tag{37}$$

and note that both $Q^{re}(\omega)$ and $Q^{im}(\omega)$ are symmetric matrices (but not necessarily positive definite). As noted in Proposition 2, the optimal strategy is to select the scalings to be the same for both sequences. Now, write (36) as

$$\rho_{12}^2(\omega;\boldsymbol{b}) = [\boldsymbol{b}'Q^{re}(\omega)\boldsymbol{b}]^2 + [\boldsymbol{b}'Q^{im}(\omega)\boldsymbol{b}]^2. \tag{38}$$

Given consistent spectral estimates $\widehat{f}_{ij}(\omega)$, we can estimate $f(\omega)$ by $\widehat{f}(\omega) = 1/2[\widehat{f}_{11}(\omega) + \widehat{f}_{22}(\omega)]$ so that consistent estimates of $Q^{re}(\omega)$ and $Q^{im}(\omega)$ are, respectively,

$$\widehat{Q}^{re}(\omega) = [\widehat{f}_{11}^{re}(\omega) + \widehat{f}_{22}^{re}(\omega)]^{-1/2}\,[\widehat{f}_{12}^{re}(\omega) + \widehat{f}_{21}^{re}(\omega)]\,[\widehat{f}_{11}^{re}(\omega) + \widehat{f}_{22}^{re}(\omega)]^{-1/2}, \tag{39}$$

$$\widehat{Q}^{im}(\omega) = [\widehat{f}_{11}^{re}(\omega) + \widehat{f}_{22}^{re}(\omega)]^{-1/2}\,[\widehat{f}_{12}^{im}(\omega) - \widehat{f}_{21}^{im}(\omega)]\,[\widehat{f}_{11}^{re}(\omega) + \widehat{f}_{22}^{re}(\omega)]^{-1/2}. \tag{40}$$

The estimated squared coherency can be maximized via Proposition 2 and the optimal scaling vector at any particular frequency, $\widehat{\boldsymbol{\beta}}(\omega)$, is taken proportional to $\widehat{f}^{re}(\omega)^{-1/2}\widehat{\boldsymbol{b}}(\omega)$, where $\widehat{\boldsymbol{b}}(\omega)$ is the maximizing vector.

4.3. Data analysis

In Fig. 4, it is seen that, although a cycle of 1/3 could be found in most of the gene, the last 1000 bp appeared to contain no cyclic behavior and might be considered to be non-coding. Herpesvirus saimiri (HVS) also contains a gene labeled BNRF1. The spectral envelopes of the last 1000 bp of the BNRF1 gene in HVS and EBV are shown in Fig. 7. Unlike EBV-BNRF1, the spectral envelope for HVS-BNRF1 has considerable power at frequency 1/3 in the final 1000 bp. It is of interest to know if the two genes match in the final 1000 bp, even though no evidence exists that the last part of EBV-BNRF1 is actually coding. Figure 8 compares the local and global alignment methods and we note significant coherency between the sequences near the one-third frequency, at least. Thus, based on the local model, we are lead to conclude that there is a significant match between the two genes in the final 1000 bp. The estimated optimal common scaling at the one-third frequency for the local model was $\mathrm{A} = 59.4, \mathrm{C} = 0.8, \mathrm{G} = 64.9, \mathrm{T} = 0$ (the global model had $\mathrm{A} = 60.8, \mathrm{C} = 5.6, \mathrm{G} = 67.1, \mathrm{T} = 0$), which indicates that the match is in the purine–pyrimidine (A = G, C = T) alphabet.

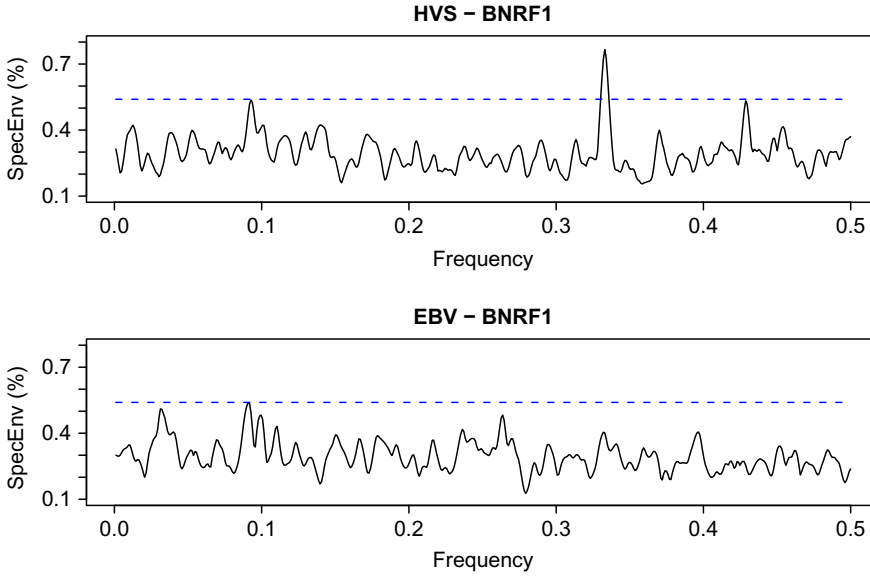

Fig. 7. Spectral envelopes of the last 1000 bp of the BNRF1 gene in the Herpesvirus saimiri (HVS) and in the Epstein–Barr virus (EBV). The horizontal dashed line indicates a pointwise 0.001 null significance threshold.

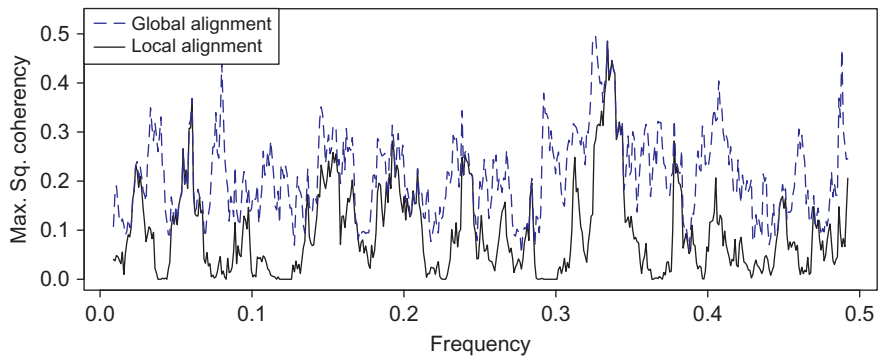

Fig. 8. Maximum squared coherency between part of EBV-BNRF and HVS-BNRF using two models, the local model, (26), and the global model, (33). The pointwise 0.0001 null significance threshold for the local model is 41.2%.

Appendix: Principal component and canonical correlation analysis for time series

As previously mentioned, the theory and methods associated with the spectral envelope and maximal coherency presented here are closely related to the theory and methods associated with principal component analysis and canonical correlation analysis for time series. In this appendix, we summarize the techniques so that the connection between the concepts is evident.

A.1. Principal components

For the case of principal component analysis for time series, suppose we have a zero mean, $p \times 1$, stationary vector process $\boldsymbol{X}_t$ that has a $p \times p$ spectral density matrix given by $f_{xx}(\omega)$. Recall $f_{xx}(\omega)$ is a complex-valued, nonnegative-definite, Hermitian matrix. Using the analogy of classical principal components, suppose for a fixed value of ω, we want to find a complex-valued univariate process $Y_t(\omega) = \boldsymbol{c}(\omega)^* \boldsymbol{X}_t$, where $\boldsymbol{c}(\omega)$ is complex, such that the spectral density of $Y_t(\omega)$ is maximized at frequency ω, and $\boldsymbol{c}(\omega)$ is of unit length, $\boldsymbol{c}(\omega)^* \boldsymbol{c}(\omega) = 1$. Because, at frequency ω, the spectral density of $Y_t(\omega)$ is $f_y(\omega) = \boldsymbol{c}(\omega)^* f_{xx}(\omega) \boldsymbol{c}(\omega)$, the problem can be restated as, find complex vector $\boldsymbol{c}(\omega)$ such that

$$\max_{\boldsymbol{c}(\omega) \neq \boldsymbol{0}} \frac{\boldsymbol{c}(\omega)^* f_{xx}(\omega) \boldsymbol{c}(\omega)}{\boldsymbol{c}(\omega)^* \boldsymbol{c}(\omega)}. \tag{A.1}$$

Let $\{(\lambda_1(\omega), \boldsymbol{e}_1(\omega)), \ldots, (\lambda_p(\omega), \boldsymbol{e}_p(\omega))\}$ denote the eigenvalue–eigenvector pairs of $f_{xx}(\omega)$, where $\lambda_1(\omega) \geq \lambda_2(\omega) \geq \cdots \geq \lambda_p(\omega) \geq 0$, and the eigenvectors are of unit length. We note that the eigenvalues of a Hermitian matrix are real. The solution to (A.1) is to choose $\boldsymbol{c}(\omega) = \boldsymbol{e}_1(\omega)$, in which case the desired linear combination is $Y_t(\omega) = \boldsymbol{e}_1(\omega)^* \boldsymbol{X}_t$. For this choice,

$$\max_{\boldsymbol{c}(\omega) \neq \boldsymbol{0}} \frac{\boldsymbol{c}(\omega)^* f_{xx}(\omega) \boldsymbol{c}(\omega)}{\boldsymbol{c}(\omega)^* \boldsymbol{c}(\omega)} = \frac{\boldsymbol{e}_1(\omega)^* f_{xx}(\omega) \boldsymbol{e}_1(\omega)}{\boldsymbol{e}_1(\omega)^* \boldsymbol{e}_1(\omega)} = \lambda_1(\omega). \tag{A.2}$$

This process may be repeated for any frequency ω, and the complex-valued process, $Y_{t1}(\omega) = \boldsymbol{e}_1(\omega)^* \boldsymbol{X}_t$, is called the first principal component at frequency ω. The kth principal component at frequency ω, for $k = 1, 2, \ldots, p$, is the complex-valued time series $Y_{tk}(\omega) = \boldsymbol{e}_k(\omega)^* \boldsymbol{X}_t$, in analogy to the classical case. In this case, the spectral density of $Y_{tk}(\omega)$ at frequency ω is $f_{y_k}(\omega) = \boldsymbol{e}_k(\omega)^* f_{xx}(\omega) \boldsymbol{e}_k(\omega) = \lambda_k(\omega)$.

The previous development of spectral domain principal components is related to the spectral envelope methodology as discussed around Eq. (5). In particular, the spectral envelope is a principal component analysis on the real part of $f_{xx}(\omega)$. Hence, the difference between spectral domain principal component analysis and the spectral envelope is that, for the spectral envelope, the $\boldsymbol{c}(\omega)$ are restricted to be real. If, in the development of the spectral envelope, we allowed for complex scalings, the two methods would be identical.

Another way to motivate the use of principal components in the frequency domain was given by Brillinger (1981, Chapter 9). Although the technique appears to be different, it leads to the same analysis. In this case, we suppose we have a stationary, p-dimensional, vector-valued process $\boldsymbol{X}_t$ and we are only able to keep a univariate process Y_t such that, when needed, we may reconstruct the vector-valued process, $\boldsymbol{X}_t$, according to an optimality criterion. Specifically, we suppose we want to approximate a mean-zero, stationary, vector-valued time series, $\boldsymbol{X}_t$, with spectral matrix $f_{xx}(\omega)$, by a univariate process Y_t defined by

$$Y_t = \sum_{j=-\infty}^{\infty} \boldsymbol{c}_{t-j}^* \boldsymbol{X}_j, \tag{A.3}$$

where $\{\boldsymbol{c}_j\}$ is a $p \times 1$ vector-valued filter, such that $\{\boldsymbol{c}_j\}$ is absolutely summable; that is, $\sum_{j=-\infty}^{\infty} |\boldsymbol{c}_j| < \infty$. The approximation is accomplished, so the reconstruction of $\boldsymbol{X}_t$ from $\boldsymbol{y}_t$, say,

$$\widehat{\boldsymbol{X}}_t = \sum_{j=-\infty}^{\infty} \boldsymbol{b}_{t-j} Y_j, \tag{A.4}$$

where $\{\boldsymbol{b}_j\}$ is an absolutely summable $p \times 1$ filter, is such that the mean square approximation error

$$E\{(\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t)^*(\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t)\} \tag{A.5}$$

is minimized.

Let $\boldsymbol{b}(\omega)$ and $\boldsymbol{c}(\omega)$ be the transforms of $\{\boldsymbol{b}_j\}$ and $\{\boldsymbol{c}_j\}$, respectively. For example,

$$\boldsymbol{c}(\omega) = \sum_{j=-\infty}^{\infty} \boldsymbol{c}_j \exp(-2\pi i j \omega), \tag{A.6}$$

and, consequently,

$$\boldsymbol{c}_j = \int_{-1/2}^{1/2} \boldsymbol{c}(\omega) \exp(2\pi i j \omega) d\omega. \tag{A.7}$$

Brillinger (1981, Theorem 9.3.1) shows that the solution to the problem is to choose $\boldsymbol{c}(\omega)$ to satisfy (A.1) and to set $\boldsymbol{b}(\omega) = \overline{\boldsymbol{c}(\omega)}$. This is precisely the previous problem, with the solution given by (A.2). That is, we choose $\boldsymbol{c}(\omega) = \boldsymbol{e}_1(\omega)$ and $\boldsymbol{b}(\omega) = \overline{\boldsymbol{e}_1(\omega)}$; the filter values can be obtained via the inversion formula given by (A.7). Using these results, in view of (A.3), we may form the first principal component series, say Y_{t1}.

This technique may be extended by requesting another series, say, Y_{t2}, for approximating $\boldsymbol{X}_t$ with respect to minimum mean square error, but where the coherency between Y_{t2} and Y_{t1} is zero. In this case, we choose $\boldsymbol{c}(\omega) = \boldsymbol{e}_2(\omega)$. Continuing this way, we can obtain the first $q \leq p$ principal components series, say, $\boldsymbol{Y}_t = (Y_{t1}, \ldots, Y_{tq})'$, having spectral density $f_{yy}(\omega) = \text{diag}\{\lambda_1(\omega), \ldots, \lambda_q(\omega)\}$. The series Y_{tk} is the kth principal component series.

A.2. Canonical correlation

In Section 4, below equation (25), we discuss the relationship between the problem of matching DNA sequences and canonical correlation analysis of time series. Here, we elaborate on the details of the relationship. Suppose we have stationary, mean-zero, $k_1 \times 1$ time series $\boldsymbol{X}_{t1}$ and $k_2 \times 1$ time series $\boldsymbol{X}_{t2}$, with respective nonsingular spectral density matrices, $f_{11}(\omega)$ and $f_{22}(\omega)$. The cross-spectral matrix between $\boldsymbol{X}_{t1}$ and $\boldsymbol{X}_{t2}$ is the $k_1 \times k_2$ matrix containing the cross-spectra between the components of $\boldsymbol{X}_{t1}$ and $\boldsymbol{X}_{t2}$. We will denote this matrix by $f_{12}(\omega)$ and note $f_{21}(\omega) = f_{12}^*(\omega)$.

In analogy to classical canonical correlations, we suppose we want to find, at a specific frequency ω, the complex linear combinations, $U_t(\omega) = \boldsymbol{\alpha}^* \boldsymbol{X}_{t1}$, and $V_t(\omega) = \boldsymbol{\beta}^* \boldsymbol{X}_{t1}$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $k_1 \times 1$ and $k_2 \times 1$ complex vectors, respectively, such that the squared coherency $\rho^2_{uv}(\omega)$ between $U_t(\omega)$ and $V_t(\omega)$ is maximum. Noting the spectral density of U_t at ω is $f_{uu}(\omega) = \boldsymbol{\alpha}^* f_{11}(\omega)\boldsymbol{\alpha}$, the spectral density of $V_t(\omega)$ at ω is $f_{vv}(\omega) = \boldsymbol{\beta}^* f_{11}(\omega)\boldsymbol{\beta}$, and the cross-spectrum between $U_t(\omega)$ and $V_t(\omega)$ is $f_{uv}(\omega) = \boldsymbol{\alpha}^* f_{12}(\omega)\boldsymbol{\beta}$, we have

$$\rho^2_{uv}(\omega) = \frac{|f_{uv}(\omega)|^2}{f_{uu}(\omega)\, f_{vv}(\omega)} = \frac{|\boldsymbol{\alpha}^* f_{12}(\omega)\boldsymbol{\beta}|^2}{[\boldsymbol{\alpha}^* f_{11}(\omega)\boldsymbol{\alpha}]\,[\boldsymbol{\beta}^* f_{22}(\omega)\boldsymbol{\beta}]}. \tag{A.8}$$

Calling the solutions $\boldsymbol{\alpha} = \boldsymbol{\alpha}_1(\omega)$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_1(\omega)$, we choose $\boldsymbol{\alpha}_1(\omega)$ to be proportional to the first eigenvector of $f_{11}^{-1/2}(\omega) f_{12}(\omega) f_{22}^{-1}(\omega) f_{21}(\omega) f_{11}^{-1/2}(\omega)$ and choose $\boldsymbol{\beta}_1(\omega)$ to be proportional to the first eigenvector of $f_{22}^{-1/2}(\omega) f_{21}(\omega) f_{11}^{-1}(\omega) f_{12}(\omega) f_{22}^{-1/2}(\omega)$. The maximum squared coherency at ω is the largest eigenvalue, $\lambda_1^2(\omega)$, of $f_{11}^{-1}(\omega) f_{12}(\omega) f_{22}^{-1}(\omega) f_{21}(\omega)$. Typically, $\boldsymbol{\alpha}_1(\omega)$ and $\boldsymbol{\beta}_1(\omega)$ are subject to the constraints

$$\boldsymbol{\alpha}_1(\omega)^* f_{11}(\omega)\boldsymbol{\alpha}_1(\omega) = 1 \quad \text{and} \quad \boldsymbol{\beta}_1(\omega)^* f_{22}(\omega)\boldsymbol{\beta}_1(\omega) = 1,$$

respectively. In this case,

$$\max \rho^2_{uv}(\omega) = |\boldsymbol{\alpha}_1^*(\omega) f_{12}(\omega)\boldsymbol{\beta}_1(\omega)|^2 = \lambda_1^2(\omega). \tag{A.9}$$

The other canonical series are selected in an obvious fashion by analogy to the classical case.

As in principal components, another view of canonical analysis exists, and this is the approach taken by Brillinger (1981, Chapter 10). Here, consider $k_i \times 1$ linear filters $\{\boldsymbol{b}_{ti}\}$ such that $\sum_t |\boldsymbol{b}_{ti}| < \infty$, $i = 1, 2$. The real-valued univariate series

$$U_t = \sum_{j=-\infty}^{\infty} \boldsymbol{b}^*_{t-j,1} \boldsymbol{X}_{j1} \quad \text{and} \quad V_t = \sum_{j=-\infty}^{\infty} \boldsymbol{b}^*_{t-j,2} \boldsymbol{X}_{j2},$$

having maximum squared coherency, $\rho^2_{uv}(\omega)$, at each ω, and subject to the constraints

$$\boldsymbol{b}_i^*(\omega) f_{ii}(\omega)\boldsymbol{b}_i(\omega) = 1,$$

for $i = 1, 2$, where $\boldsymbol{b}_i(\omega)$ is the transform of $\{\boldsymbol{b}_{ti}\}$, are given by finding the largest scalar $\lambda(\omega)$ such that

$$f_{11}(\omega)^{-1/2} f_{12}(\omega) f_{22}(\omega)^{-1} f_{21}(\omega) f_{11}(\omega)^{-1/2} \boldsymbol{\alpha}(\omega) = \lambda^2(\omega)\boldsymbol{\alpha}(\omega) \tag{A.10}$$

and

$$f_{22}(\omega)^{-1/2} f_{21}(\omega) f_{11}(\omega)^{-1} f_{12}(\omega) f_{22}(\omega)^{-1/2} \boldsymbol{\beta}(\omega) = \lambda^2(\omega)\boldsymbol{\beta}(\omega). \tag{A.11}$$

The maximum squared coherency achieved between U_t and V_t is $\lambda^2(\omega)$, and $\boldsymbol{b}_1(\omega)$ and $\boldsymbol{b}_2(\omega)$ are taken proportional to first eigenvectors of

$$f_{11}^{-1/2}(\omega) f_{12}(\omega) f_{22}^{-1}(\omega) f_{21}(\omega) f_{11}^{-1/2}(\omega)$$

and of

$$f_{22}^{-1/2}(\omega) f_{21}(\omega) f_{11}^{-1}(\omega) f_{12}(\omega) f_{22}^{-1/2}(\omega),$$

respectively. The required filters can be obtained by inverting $\boldsymbol{b}_1(\omega)$ and $\boldsymbol{b}_2(\omega)$ using a relationship such as (A.7). Again, the other canonical variate series are obtained in an obvious way, and estimation proceeds by replacing the spectra $f_{ij}(\omega)$ by their respective estimates $\widehat{f_{ij}}(\omega)$, for $i, j = 1, 2$.

Acknowledgment

This work was supported, in part, by grant DMS-0805050 from the National Science Foundation.

References

Adak, S., 1998. Time dependent spectral analysis of nonstationary time series. J. Am. Stat. Assoc. 93, 1488–1501.

Bernardi, G., Bernardi, G., 1985. Codon usage and genome composition. J. Mol. Evol., 22, 363–365.

Bina, M., 1994. Periodicity of dinucleotides in nucleosomes derived from simian virus 40 chromatin. J. Mol. Biol. 235, 198–208.

Blaisdell, B.E., 1983. Choice of base at silent codon site 3 is not selectively neutral in eucaryotic structural genes: it maintains excess short runs of weak and strong hydrogen bonding bases. J. Mol. Evol. 19, 226–236.

Blackman, R.B., Tukey, J.W., 1959. The Measurement of Power Spectra from the Point of View of Communications Engineering. Dover, New York.

Brillinger, D.R., 1981. Time Series: Data Analysis and Theory, second ed. Holden-Day.

Brillinger, D.R., 2001. Time Series: Data Analysis and Theory, Society for Industrial and Applied Mathematics.

Buckingham, R.H., 1990. Codon context. Experientia 46, 1126–1133.

Cornette, J.L., Cease, K.B., Margaht, H., Spouge, J.L., Berzofsky, J.A., DeLisi, C., 1987 Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. J. Mol. Biol. 195, 659–685.

Curran, J.F., Gross, B.L., 1994. Evidence that GHN phase bias does not constitute a framing code. J. Mol. Biol. 235, 389–395.

Dahlhaus, R., 1997. Fitting time series models to nonstationary processes. Ann. Stat., 25, 1–37.

Drew, H.R., Calladine, C.R., 1987. Sequence-specific positioning of core histones on an 860 base-pair DNA: Experiment and theory. J. Mol. Biol. 195, 143–173.

Eisenberg, D., Weiss, R.M., Terwillger, T.C., 1994. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc. Natl. Acad. Sci. 81, 140–144.

Ioshikhes, I., Bolshoy, A., Trifonov, E.N., 1992. Preferred positions of AA and TT dinucleotides in aligned nucleosomal DNA sequences. J. Biomol. Struct. Dyn. 9, 1111–1117.

Jeong, H., 2011. The Spectral Analysis of Nonstationary Categorical Time Series Using Local Spectral Envelope. Ph.D. Dissertation, University of Pittsburgh.

Komberg, R.D., 1974. Chromatin structure: A repeating unit of histones and DNA. Science, 184, 868–871.

Lagunez-Otero, J., Trifonov, E.N., 1992. mRNA periodical infrastructure complementary to the proof-reading site in the ribosome. J. Biomol. Struct. Dyn. 10, 455–464.
Maddox, J., 1992. Long-range correlations within DNA. Nature, 358, 103.
McLachlan, A.D., Stewart, M., 1976. The 14-fold periodicity in alpha-tropomyosin and the interaction with actin. J. Mol. Biol. 103, 271–298.
Mengeritsky, G., Trifonov, E.N.,1983. Nucleotide sequence-directed mapping of the nucleosomes. Nucleic Acids Res. 11, 3833–3851.
Muyldermans, S., Travers, A.A., 1994 DNA sequence organization in chromatosomes. J. Mol. Biol. 235, 855–870.
Piña, B., Barettino D., Truss, M., Beato, M., 1990. Structural features of a regulatory nucleosome. J. Mol. Biol. 216, 975–990.
Riordan, J. R., Rommens, J.M., Kerem, B.S., Alon, N., Rozmahel, R., Grzelczak, Z., et al., 1989. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science, 245, 1066–1073.
Rosen, O., Stoffer, D.S., 2007. Automatic estimation of multivariate spectra via smoothing splines. Biometrika, 94, 335–345.
Satchwell, S.C., Drew, H.R., Travers, A.A., 1986. Sequence periodicities in chicken nucleosome core DNA. J. Mol. Biol. 191, 659–675.
Schachtel, G.A., Bucher, P., Mocarski, E.S., Blaisdell, B.E., Karlin, S., 1991. Evidence for selective evolution in codon usage in conserved amino acid segments of human alphaherpesvirus proteins. J. Mol. Evol. 33, 483–494.
Shepherd, J.C.W., 1984. Fossil remnants of a primeval genetic code in all forms of life? Trends Biochem. Sci. 9, 8–10.
Shrader, T.E., Crothers, D.M., 1990. Effects of DNA sequence and histone-histone interactions on nucleosome placement. J. Mol. Biol. 216, 69–84.
Shumway, R.H., Stoffer, D.S., 2011. Time Series Analysis and Its Applications: With R Examples, third ed. Springer, New York.
Simpson, R.T., 1990. Nucleosome positioning can affect the function of a cis-acting DNA element in vivo. Nature, 343, 387–389.
Stoffer, D.S. (1987). Walsh-Fourier analysis of discrete-valued time series. J. Time Series Anal., 8, 449–467.
Stoffer, D.S., 1991. Walsh-Fourier analysis and its statistical applications (with discussion). J. Am. Stat. Assoc. 86, 462–483.
Stoffer, D.S., Ombao, H., Tyler, D.E., 2002. Evolutionary spectral envelope: An approach using tree-based adaptive segmentation. Ann. Inst. Stat. Math. 54, 201–223.
Stoffer, D.S., Tyler, D.E., 1998. Matching sequences: Cross spectral analysis of categorical time series. Biometrika, 85, 201–213.
Stoffer, D.S., Tyler, D.E., McDougall, A.J., 1993a. Spectral analysis for categorical time series: Scaling and the spectral envelope. Biometrika, 80, 611–622.
Stoffer, D.S., Tyler, D.E., McDougall, A.J., Schachtel, G.A., 1993b. Spectral analysis of DNA sequences (with discussion). Bull. Int. Stat. Inst. Bk 1, 345–361; Bk 4, 63–69.
Sueoka, N., 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. 85, 2653–2657.
Tavaré, S., Giddings, B.W., 1989. Some statistical aspects of the primary structure of nucleotide sequences. In Waterman M.S. (Ed), Mathematical Methods for DNA Sequences. CRC Press, Boca Raton, Florida, pp. 117–131.
Trifonov, E.N., 1987. Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences. J. Mol. Biol. 194, 643–652.
Trifonov, E.N., Sussman, J.L., 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc. Natl. Acad. Sci. 77, 3816–3820.
Turnell, W.G., Satchwell, S.C., Travers, A.A., 1988. A decapeptide motif for binding to the minor groove of DNA: A proposal. Febs Lett. 232, 263–268.
Uberbacher, E.C., Harp, J.M., Bunick, G.J., 1988. DNA sequence patterns in precisely positioned riucleosomes. J. Biomol. Struct. Dyn. 6, 105–120.
Viari, A., Soldano, H., Ollivier, E., 1990. A scale-independent signal processing method for sequence analysis. Comput. Appl. Biosci. 6, 71–80.

Waterman, M. S., Vingron, M., 1994. Sequence comparison significance and Poisson approximation. Stat. Sci. 9, 367–381.

Whisenant E.C., Rasheed, B.K., Ostrer, H., Bhatnagar, Y.M., 1991. Evolution and sequence analysis of a human Y-chromosomal DNA fragment. J. Mol. Evol. 33, 133–141.

Wong, J.T., Cedergren, R., 1986. Natural selection versus primitive gene structure as determinant of codon usage. Eur. J. Biochem. 169, 175–180.

Yarus, M., Folley, L.S., 1985. Sense codons are found in specific contexts. J. Mol. Biol. 1985 Apr 20, 182(4):529–40.

Zhurkin, V.B., 1983. Specific alignment of nucleosomes on DNA correlates with periodic distribution of purine-pyrimidine and pyrimidine-purine dimers. FEBS Lett. 158, 293–297.

Zhurkin, V.B., 1985. Sequence-dependent bending of DNA and phasing of nucleosomes. J. Biomol. Struct. Dyn. 2, 785–804.

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00011-9

Spatial Time Series Modeling for fMRI Data Analysis in Neurosciences

Tohru Ozaki
IDAC, Tohoku University, Sendai, Japan

Abstract

The statistical analysis of spatial and temporal data is discussed from the viewpoint of an fMRI connectivity study. The limitations of the well-known SPM method for the characterization of fMRI connectivity study are pointed out. The use of an innovation approach with NN-ARX is suggested to overcome the limitations of the SPM modeling. The maximum likelihood method is presented for the NN-ARX model estimation. The exploratory use of innovations for the identification of brain connectivity between remote voxels is discussed.

Keywords: spatial temporal correlations, spatial stochastic process, innovation approach, state space modeling, fMRI data, SPM, NN-ARX model, brain connectivity, causality.

1. Introduction

The main objective of this chapter is to show that the innovation approach, developed for time series analysis, is useful for the characterization of the spatial and temporal dynamic structure of the "very high" dimensional spatial temporal data such as fMRI (functional magneto-resonance imaging) in neuroscience.

fMRI is the measurement of BOLD (Blood-Oxygen-Level Dependence) signal, which is a hemodynamic response related to neural activity in the brain or spinal cord of humans or other animals. fMRI data are obtained as a time series at each voxel inside the brain; thus, the fMRI data are typically a high-dimensional ($64 \times 64 \times 36$) time series observed from one experiment on a subject usually under some properly designed

stimulus. In the live human brain, more neuronal activity requires more glucose and oxygen to be delivered rapidly through the blood stream. It results in a surplus of oxyhemoglobin in the veins of the area and distinguishable change of the local ratio of oxyhemoglobin to deoxyhemoglobin. The higher BOLD signal intensities arise from increases in the concentration of oxygenated hemoglobin since the magnetic susceptibility of blood now more closely matches with that of the tissue. Here, the fMRI data are obtained by collecting data in an MRI (Magneto Resonance Imaging) scanner with sequence parameters sensitive to changes in magnetic susceptibility, and one can assess changes in BOLD contrast (Buxton, 2002). Thus, the fMRI is a type of specialized Magneto Resonance Imaging (MRI) scan. The problem with original MRI scan technology was that although it provides a detailed assessment of the physical appearance, water content, and many kinds of subtle derangements of structure of the brain (such as inflammation or bleeding), it fails to provide information about the metabolism of the brain (i.e., how actively it is "functioning") at the time of imaging. Therefore, a distinction is made between "MRI imaging" and "functional MRI imaging" (fMRI), where MRI provides only structural information on the brain while fMRI yields both structural and "dynamic" (functional) data.

The dynamic neural association between spatially remote distinct brain regions gives us the key to understand human brain functions, and fMRI data are becoming a more and more common tool for brain connectivity studies. According to Friston and Buckel (2004), the definition of functional connectivity is "correlation between spatially remote neurophysiological events." fMRI (BOLD signal) time series data are a kind of spatial temporal data, and what is needed here is an efficient statistical method for the estimation of "spatial temporal correlation structure" from the fMRI time series data, which usually consists of about 147,000 channel time series measured at $64 \times 64 \times 36$ grid points in the brain.

2. A traditional approach: Spatial and temporal covariance functions

A traditional approach to the statistical characterization of spatial and temporal data is to estimate the covariance and correlation structure of the data. The correlation structure of a spatial temporal process behind the fMRI data, $\{x_t^{(i,j,k)}(t = 1, 2, \ldots, N)\}$ with $\{(i, j, k) \in (64 \times 64 \times 36)\}$, may be derived from the spatial and temporal covariance function,

$$c_x(h, k) = \mathrm{cov}\left\{\tilde{x}_{t+k}^{(s+h)}, \tilde{x}_t^{(s)}\right\}$$

where $\tilde{x}_t = x_t - E[x_t]$.

In much of the literature (Cressie (1993) for example), it is assumed that the spatial temporal covariance function has the following product representation,

$$c_x(h, k) = c_x^{(1)}(h)c_x^{(2)}(k)$$

with a purely spatial component $c_x^{(1)}(h)$ and a purely temporal component $c_x^{(2)}(k)$.

However, it must be noted that this assumption, called "separability," is purely artificial and there is no physical or physiological reason supporting this assumption. Although the spatial covariance function of a spatial temporal process $x_t^{(s)}$ may be formally defined (Whittle, 1962), under the spatial stationarity assumption, as

$$c_x^{(1)}(h) = \text{cov}\left\{\tilde{x}_t^{(s+h)}, \tilde{x}_t^{(s)}\right\},$$

it is not so useful a tool as the temporal covariance function $E[\tilde{x}_t^{(s)}\tilde{x}_{t+\tau}^{(s)}] = R_{xx}^{(s)}(\tau)$ defined in neuroscience at each voxel $s = (i, j, k)$. A large spatial correlation between two cortexes, for example the primary visual cortex $V1$ at (i, j, k) and the visual cortex $V5$ at $(i + u1, j + u2, k + u3)$, does not imply similar large spatial correlations for any other voxels (x, y, z) and $(x + u1, y + u2, z + u3)$ with the same spatial shift determined by $u1$, $u2$, and $u3$. This kind of spatial stationarity assumption may be suitable for the analysis of agricultural yield in the plane where the space is isotropic and the specific position does not matter (Whittle, 1962), but it is surely not suitable for the human brain functioning.

Unlike the spatial covariance function, the temporal covariance function is a reasonable measure for the characterization of the spatial process underlying the fMRI time series. When the brain is under control for some time interval, with or without a stimulus, the measured fMRI time series may be considered to be temporally stationary. Under the assumption of temporal stationarity, we can define the temporal covariance functions at each voxel (i, j, k) as

$$R_{xx}^{(i,j,k)}(\tau) = E\left[x_t^{(i,j,k)} x_{t+\tau}^{(i,j,k)}\right] \tag{1}$$

When fMRI data are measured under the on–off block-designed external stimulus, the temporal covariance function needs to be redefined carefully. If we could assume that the stimulus affects the BOLD signal continuously and the effect of stimuli continues beyond the resting period during the experiment, the definition of (1) may still be valid. However, if the effect of the stimulus on the BOLD signal is significantly large compared with the period when stimulus is off, the relation between the connectivity and the temporal covariance function is not so simple. The temporal covariance function under on-mode and off-mode has to be distinguished through elaborate modeling in time domain (see Yamashita et al. (2005)).

3. SPM and the implied determinism

In neuroscience, one of the most commonly used methods for the statistical analysis of spatial and temporal structure of fMRI data is not the classical spatial temporal covariance function approach but a statistical method called SPM (Friston et al., 1995) and its later versions. The SPM approach brings together two well-established bodies of theory (the general linear model and the theory of Gaussian fields) to provide a complete and simple framework for the statistical analysis of imaging data. Indeed, the spatial

information in fMRI is well exploited by sophisticated spatial statistical methods in SPM. However, it must be noted that, in SPM, the temporal correlation information is not fully exploited.

The fMRI data $x_t^{(v)}$ (after going through necessary preprocessing such as normalization and removing artifacts) are characterized in SPM as

$$x_t^{(v)} = \sum_{k=0}^{T} h_k s_{t-k} + \xi_t^{(v)} \tag{2}$$

Here, the noise $\xi_t^{(v)}$ is an observation error and does not affect the future trajectory of $x_t^{(v)}$. The response function h_k is given by a combination of Gamma functions (see Lange and Zeger (1997) and Worsley et al. (2002)), for example, $t^r e^{-\lambda t}$. The assumed parametric response function is subsampled at n scan acquisition times $t_1, t_2, \ldots, t_n$ to give the response $y_i = y(t_i)$ at scan i. Then, the observed fMRI data x_i may be explained using an observation error ξ_i as

$$x_i = y_i \beta + \xi_i,$$

where the parameters in the model are estimated by the least squares method. Then (2) is equivalent to

$$\begin{aligned} y_t^{(v)} &= \sum h_k s_{t-k} \\ x_t^{(v)} &= y_t^{(v)} + \xi_t^{(v)} \end{aligned} \tag{3}$$

The model implied by (3) is a deterministic process $y_t^{(v)}$, driven by an exogenous process $s(t)$, measured by $x_t^{(v)}$ with an additive observation noise $\xi_t^{(v)}$, that is,

$$\begin{aligned} \frac{dz^{(v)}(t)}{dt} &= Az^{(v)}(t) + Bs(t) \\ x_t^{(v)} &= Cz^{(v)}(t) + \xi_t^{(v)} \end{aligned} \tag{4}$$

For example, the deterministic system implied by the response function $t^{k-1}e^{-\lambda t}$ is given by (4), where the $k \times k$ matrix A, k-dimensional vectors B and C are given by,

$$A = \begin{pmatrix} -\lambda & 0 & \ldots & 0 & 0 \\ 1 & -\lambda & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & -\lambda & 0 \\ 0 & 0 & \ldots & 1 & -\lambda \end{pmatrix}, \quad B = \begin{pmatrix} b^{(v)} \\ 0 \\ \ldots \\ 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad C = (0\ 0\ \ldots\ 0\ 1)$$

If we need a more general response function, we need to use a more general state space model. For example, in order to have the following impulse response function,

$$h(t) = h_1 e^{-\lambda_1 t} + h_2 t e^{-\lambda_2 t} + \cdots + h_k t^k e^{-\lambda_k t},$$

we need to have the following state space model,

$$\frac{d\begin{pmatrix} z_1^{(1)}(t) \\ z_1^{(2)}(t) \\ z_2^{(2)}(t) \\ \cdots \\ z_1^{(k)}(t) \\ z_2^{(k)}(t) \\ \cdots \\ z_{k-1}^{(k)}(t) \\ z_k^{(k)}(t) \end{pmatrix}}{dt} = \begin{pmatrix} -\lambda_1 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & 0 \\ 0 & -\lambda_2 & 0 & \ldots & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & -\lambda_2 & \ldots & 0 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & -\lambda_k & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & \ldots & 1 & -\lambda_k & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 & 0 & \ldots & -\lambda_k & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 1 & -\lambda_k \end{pmatrix}$$

$$\times \begin{pmatrix} z_1^{(1)}(t) \\ z_1^{(2)}(t) \\ z_2^{(2)}(t) \\ \cdots \\ z_1^{(k)}(t) \\ z_2^{(k)}(t) \\ \cdots \\ z_{k-1}^{(k)}(t) \\ z_k^{(k)}(t) \end{pmatrix} + \begin{pmatrix} b_1^{(v)} \\ b_2^{(v)} \\ 0 \\ \cdots \\ b_k^{(v)} \\ 0 \\ \cdots \\ 0 \\ 0 \end{pmatrix} s(t)$$

$$x_t^{(v)} = z_1^{(1)}(t) + z_2^{(2)}(t) + \cdots + z_k^{(k)}(t) + \xi_t^{(v)}$$

The discrete time version of the model (4) is written as,

$$\begin{aligned} Z_t^{(v)} &= AZ_{t-1}^{(v)} + B^{(v)}s_{t-1} \\ x_t^{(v)} &= CZ_t^{(v)} + \xi_t^{(v)} \end{aligned} \tag{5}$$

Here, the dimension of the state $Z_t(v)$ is $K = k(k+1)/2$. What this deterministic model (4) or (5) implies is a very strong assumption that the future value of the BOLD signal $y_{t+\tau}^{(v)}(\tau > 0)$ is exactly predicted by the initial state $Z_0^{(v)}$ and the input $s(t)(0 < t < T)$. Here, the noise $\xi_t^{(v)}$ is an observation error and never affects the future value of $Z_t^{(v)}$ or $y_t^{(v)}$. The model is equivalent to the following ARMA (K, K) model

with exogenous input (see Ozaki (1998)),

$$\begin{aligned} x_t^{(v)} + \phi_1 x_{t-1}^{(v)} + \cdots + \phi_K x_{t-K}^{(v)} &= (\xi_t + CB^{(v)} s_t) + \{C(A - \phi_1 I)B^{(v)} s_{t-1} + \phi_1 \xi_{t-1}\} \\ &\quad + \cdots + \{C(A^{K-1} + \phi_1 A^{K-2} + \cdots + \phi_{K-1} I) \\ &\quad \times B^{(v)} s_{t-K+1} + \phi_{K-1}\xi_{t-K+1}\} + \phi_K \xi_{t-K} \\ &= CB^{(v)} s_{t-1} + C(A - \phi_1 I)B^{(v)} s_{t-2} + \cdots + C(A^{K-1} \\ &\quad + \phi_1 A^{K-2} + \cdots + \phi_{K-1} I)B^{(v)} s_{t-K} + \xi_t + \phi_1 \xi_{t-1} \\ &\quad + \cdots + \phi_{K-1}\xi_{t-K+1} + \phi_K \xi_{t-K} \end{aligned}$$

We can calculate, with the assumption of the Gaussian white noise for $\xi_t^{(v)}$, (−2)log-likelihood of the ARX model (see Box and Jenkins (1970)). Note that, the AR and MA coefficients and coefficients of the input signals have strong constraint between them. This kind of constrained model is known to be very inflexible and does not produce small prediction errors, yielding larger (−2)log-likelihood and larger AIC compared with unconstrained ARMAX models. The validity of this assumption of constraining also needs to be checked by an objective statistical method based on the observed data $x_t^{(v)}$.

Note that the response function h_k $(k = 1, 2, \ldots.)$ in SPM is set as independent from the space variable v. This is also a very strong assumption, whose validity needs to be checked by an objective statistical method based on the observed fMRI data.

4. Innovation approach and the NN-ARX model

In time series analysis, a natural way of characterizing the temporal correlation structure of a stationary time series is to use a linear dynamic model such as an AR model. By identifying a suitable multivariate AR model from the observed data, we can characterize the multivariate autocovariance functions of the process behind the data.

Identification and estimation of stochastic/deterministic dynamical system models from observed time series data have been the much-studied topic since the era of N. Wiener and A.N. Kolmogorov. Many methods have been introduced and discussed since the 1930s. Among them, one approach called the "Innovation Approach" (or equivalently "Prediction Error Approach"), introduced by N. Wiener (1949), may be specially interesting and useful for applied scientists, since the guideline principle is intuitionally simple and the computational algorithm and statistical diagnostic checking is straightforward and easy to perform for practitioners.

The innovation approach suggests us to find a dynamic model yielding the smallest prediction errors from the time series data. The fMRI BOLD signal data are presented as a set of huge (typically 147,000) dimensional time series, where the prediction of each variable out of the 147,000 variables is considered. The optimal prediction of $x_t^{(i,j,k)}$ at time point $t-1$ will be $E\left[x_t^{(i,j,k)} | x_{t-1}^{(*)}\right]$ under the local Gaussian assumption. Here, $x_{t-1}^{(*)}$ denotes all the BOLD signal information at the time point $t-1$. Since neighboring voxels could contain the most useful information for the one-step ahead prediction of $x_t^{(i,j,k)}$, a natural approximate linear predictor will be a linear combination

of the neighboring voxels $x_{t-1}^{(i-1,j,k)}, x_{t-1}^{(i+1,j,k)}, \ldots x_{t-1}^{(i,j,k+1)}$ and itself $x_{t-1}^{(i,j,k)}$. Then, the prediction error may be written as

$$\varepsilon_t^{(i,j,k)} = x_t^{(i,j,k)} - \left\{ a_1^{(i,j,k)} x_{t-1}^{(i,j,k)} + b_1^{(i,j,k)} x_{t-1}^{(i-1,j,k)} + b_2^{(i,j,k)} x_{t-1}^{(i+1,j,k)} + \cdots + b_6^{(i,j,k)} x_{t-1}^{(i,j,k+1)} \right\}$$

When the fMRI of the subject is measured under some experiment with controlled stimulus inputs, the one-step ahead predictions may be significantly improved by using the information (on or off) of the stimulus in the experiment. Then, the prediction error of $x_t^{(i,j,k)}$ is,

$$\varepsilon_t^{(i,j,k)} = x_t^{(i,j,k)} - \left\{ a_1^{(i,j,k)} x_{t-1}^{(i,j,k)} + b_1^{(i,j,k)} x_{t-1}^{(i-1,j,k)} + b_2^{(i,j,k)} x_{t-1}^{(i+1,j,k)} + \cdots + b_6^{(i,j,k)} x_{t-1}^{(i,j,k+1)} + \theta_1^{(i,j,k)} s_{t-1} \right\}$$

This implies that the following spatial autoregressive type model will be a reasonable approximate dynamic initial model for an fMRI time series:

$$x_t^{(i,j,k)} = a_1^{(i,j,k)} x_{t-1}^{(i,j,k)} + b_1^{(i,j,k)} x_{t-1}^{(i-1,j,k)} + b_2^{(i,j,k)} x_{t-1}^{(i+1,j,k)} + \cdots + b_6^{(i,j,k)} x_{t-1}^{(i,j,k+1)} + \theta_1^{(i,j,k)} s_{t-1} + \varepsilon_t^{(i,j,k)}$$

The original idea of this model was introduced by the present author at the Workshop on Mathematical Methods in Brain Mapping at CRM, University of Montreal, in 2000, and has been further developed by Riera et al. (2004), where the physiological meaning of the model is clarified, and is called NN-ARX model (Nearest Neighbor AutoRegressive model with eXogenous variable). A more general NN-ARX model with higher lag orders may be written as

$$x_t^{(v)} = a_1^{(v)} x_{t-1}^{(v)} + \cdots + a_p^{(v)} x_{t-p}^{(v)} + \frac{1}{6} \left(\sum_{v' \in N(v)} b_{v'}^{(v)} x_{t-1}^{(v')} \right) + \theta_1^{(v)} s_{t-1} + \cdots + \theta_r^{(v)} s_{t-r} + \varepsilon_t^{(v)} \tag{6}$$

Here, $(\mathrm{v}) = (i,j,k), N(v) = \{(i-1,j,k), (i+1,j,k), (i,j-1,k), (i,j+1,k), (i,j,k-1), (i,j,k+1)\}$. Coefficients $a_1^{(v)}, \ldots, a_p^{(v)}, b_{v'}^{(v)}, \ldots, \theta_1^{(v)}, \ldots, \theta_r^{(v)}$ are calculated by solving the linear equation for each voxel v. Whether the system noise $\varepsilon_t^{(v)}$ is zero or not need to be checked by a statistical method. Incidentally, we note that the discrete time model (6), with $p = 1$ and $r = 1$, can be obtained by discretizing the following partial differential equation model of a spatial stochastic process with an external input $s(t)$.

$$\frac{\partial x(\xi,\eta,\varsigma,t)}{\partial t} = a(\xi,\eta,\varsigma)x + b(\xi,\eta,\varsigma)\left(\frac{\partial^2 x}{\partial \xi^2} + \frac{\partial^2 x}{\partial \eta^2} + \frac{\partial^2 x}{\partial \varsigma^2} \right) + \theta(\xi,\eta,\varsigma)s(t) + \delta W(\xi,\eta,\varsigma,t)$$

In other words, we are interpreting the fMRI data as a realization of a spatial "blurring" process driven by stimulus $s(t)$ and a spatial Gaussian white noise process $\delta W(\xi, \eta, \varsigma, t)$ (Brown et al., 2000). Here, the innovation approach with NN-ARX model performs a kind of "de-blurring" procedure in order to improve the resolution so that we may discover important temporal spatial information in the data. The theoretical foundation of the innovation (prediction error) approach to spatial stochastic processes was given by K. Ito (1984).

It must be noted that we have so far ignored possible instantaneous correlations between the noise of neighboring voxels and the noise covariance of the 147,000 dimensional AR model is assumed to be diagonal. This may not be an appropriate assumption for the NN-ARX model to be a general spatial time series model. One simple way of removing the instantaneous correlations between the neighboring voxels is to apply an instantaneous Laplacian operator L, which operates as

$$Lx_t^{(i,j,k)} = x_t^{(i,j,k)} - \frac{1}{6}\left(x_t^{(i+1,j,k)} + x_t^{(i-1,j,k)} + x_t^{(i,j+1,k)} + x_t^{(i,j-1,k)} + x_t^{(i,j,k-1)} + x_t^{(i,j,k+1)}\right)$$

for the three-dimensional case.

If we apply the Laplacian operator L to the original data before fitting the above NN-ARX model, then we have,

$$Lx_t^{(v)} = y_t^{(v)}$$

$$y_t^{(v)} = \mu_t^{(v)} + \sum_{k=1}^{r_1} \alpha_k^{(v)} y_{t-k}^{(v)} + \sum_{k=1}^{r_2} \beta_k^{(v)} \xi_{t-k}^{(v)} + \sum_{k=1}^{r_3} \gamma_k^{(v)} s_{t-k} + n_t^{(v)}$$

While the variance matrix of the noise $n_t^{(v)}$ in the transformed space is diagonal, so that $E[n_t n_t'] = \sigma_n^2 I$, the variance matrix of the noise $\varepsilon_t^{(v)} = L^{-1} n_t^{(v)}$ in the original space is nondiagonal and is given by $\Sigma_\varepsilon = \sigma_n^2 (L'L)^{-1}$. This is a simple but useful way of characterizing a spatially homogeneous instantaneous dependency between the noises of neighboring voxels and was used in EEG dynamic inverse solutions by Galka et al. (2004) and Yamashita et al. (2004). The superiority of the NN-ARX model with the Laplacian operator can be confirmed by comparing the AIC of the two models, with or without the Laplacian, fitted to the same fMRI data.

5. Likelihood and the significance of the assumptions

The (-2)log-likelihood of the NN-ARX model is given by

$$(-2)\log p\left(x_1^{(1,1,1)}, \ldots, x_1^{(64,64,36)}, \ldots, x_N^{(1,1,1)}, \ldots, x_N^{(64,64,36)} | \varphi\right)$$
$$\approx \sum_{v=(1,1,1)}^{(64,64,36)} \left[\sum_{t=p+1}^{T} \left\{\log \sigma^2_{\varepsilon_t^{(v)}} + \frac{(\varepsilon_t^{(v)})^2}{\sigma^2_{\varepsilon_t^{(v)}}}\right\}\right] + \text{Const}$$

If the data $x_t^{(v)}$ are transformed into $y_t^{(v)}$ by a Laplacian operator L, in order to remove the instantaneous correlations between neighboring voxels, the likelihood function becomes,

$$\begin{aligned}
&(-2)\log p\left(x_1^{(1,1,1)},\ldots,x_1^{(64,64,36)},\ldots,x_N^{(1,1,1)},\ldots,x_N^{(64,64,36)}|\varphi\right)\\
&\quad=(-2)\log p\left(y_1^{(1,1,1)},\ldots,y_1^{(64,64,36)},\ldots,y_N^{(1,1,1)},\ldots,y_N^{(64,64,36)}|\varphi\right)\\
&\qquad+\log\det(L^{-1})\\
&\quad\approx\sum_{v=(1,1,1)}^{(64,64,36)}\left[\sum_{t=1}^{T}\left\{\log\sigma^2_{\varepsilon_t^{(v)}}+\frac{(\varepsilon_t^{(v)})^2}{\sigma^2_{\varepsilon_t^{(v)}}}\right\}\right]+\log\det(L^{-1})+\text{Const}
\end{aligned}\tag{7}$$

$\varepsilon_t^{(v)}$ is given for a standard NN-ARX model by,

$$\varepsilon_t^{(v)}=y_t^{(v)}-\left\{a_1^{(v)}y_{t-1}^{(v)}+\frac{1}{6}\left[\sum_{v'\in N(v)}b_{v'}^{(v)}y_{t-1}^{(v')}\right]+\theta_1^{(v)}s_{t-1}\right\}\quad\text{for }\forall v$$

(−2)log-likelihood of the deterministic SPM model,

$$\begin{aligned}
Z_t^{(v)}&=AZ_{t-1}^{(v)}+B^{(v)}s_{t-1}\\
x_t^{(v)}&=CZ_t^{(v)}+\xi_t^{(v)}
\end{aligned}\tag{8}$$

is given by

$$\begin{aligned}
&(-2)\log p\left(x_1^{(1,1,1)},\ldots,x_1^{(64,64,36)},\ldots,x_N^{(1,1,1)},\ldots,x_N^{(64,64,36)}|\varphi\right)\\
&\quad=(-2)\log p\left(x_1^{(1,1,1)},\ldots,x_1^{(64,64,36)},\ldots,x_N^{(1,1,1)},\ldots,x_N^{(64,64,36)}|\varphi\right)\\
&\quad\approx\sum_{v=(1,1,1)}^{(64,64,36)}\left[\sum_{t=1}^{T}\left\{\log\sigma^2_{\varepsilon_t^{(v)}}+\frac{(\varepsilon_t^{(v)})^2}{\sigma^2_{\varepsilon_t^{(v)}}}\right\}\right]+\text{Const}
\end{aligned}$$

where $\varepsilon_t^{(v)}$ is given by the following recursive Kalman filter scheme,

$$\begin{aligned}
\varepsilon_t^{(v)}&=x_t^{(v)}-CZ_{t|t-1}^{(v)}\\
Z_{t|t-1}^{(v)}&=AZ_{t-1|t-1}^{(v)}+B^{(v)}s_{t-1}\\
Z_{t-1|t-1}^{(v)}&=Z_{t-1|t-2}^{(v)}+K_{t-1}^{(v)}\varepsilon_{t-1}^{(v)}\\
K_{t-1}^{(v)}&=P_{t-1}^{(v)}C'\left\{CP_{t-1}^{(v)}C'+\sigma^2_{\xi^{(v)}}\right\}^{-1}\\
P_{t-1}^{(v)}&=AV_{t-2}^{(v)}A'\\
V_{t-1}^{(v)}&=P_{t-1}^{(v)}-K_{t-1}^{(v)}CP_{t-1}^{(v)}
\end{aligned}\tag{9}$$

Because of the affecting mechanism of the nearest neighbors in the spatial structure, we have approximately

$$Z_{t|t-1}^{(v)} = E\left[Z_t^{(v)}|x_{t-1}^{(*)}, x_{t-2}^{(*)}, \ldots, x_1^{(*)}\right] \approx E\left[Z_t^{(v)}|x_{t-1}^{(v)}, x_{t-1}^{N(v)}, x_{t-2}^{(*)}, \ldots, x_1^{(*)}\right]$$
$$Z_{t|t}^{(v)} = E\left[Z_t^{(v)}|x_t^{(*)}, x_{t-1}^{(*)}, .x_{t-2}^{(*)}, .., x_1^{(*)}\right] \approx E\left[Z_t^{(v)}|x_t^{(v)}, x_{t-1}^{(v)}, x_{t-1}^{N(v)}, .x_{t-2}^{(*)}, \ldots, x_1^{(*)}\right].$$

Here $x_t^{(*)}$ means all the observed BOLD signals at the time point t.

5.1. *Statistical check of the assumption of determinism implied in SPM*

We can statistically check whether the deterministic model employed by SPM is justified by comparing (−2)log-likelihood of the deterministic SPM model:

$$\begin{aligned} Z_t^{(v)} &= AZ_{t-1}^{(v)} + B^{(v)}s_{t-1} \\ x_t^{(v)} &= CZ_t^{(v)} + \xi_t^{(v)} \end{aligned} \tag{10}$$

and (−2)log-likelihood of the more general stochastic model:

$$\begin{aligned} Z_t^{(v)} &= AZ_{t-1}^{(v)} + B^{(v)}s_{t-1} + Dn_t^{(v)} \\ x_t^{(v)} &= CZ_t^{(v)} + \xi_t^{(v)}. \end{aligned} \tag{11}$$

Since both models have state space representations, their (−2)log-likelihoods are calculated using the innovations obtained by the Kalman filter scheme (9). If we compare the AIC of two models, obviously the model (11) shows much smaller AIC than the model (10).

5.2. *Statistical check of the activation in each voxel*

Whether the voxel V is activated while the subject is under certain stimulus can be detected by comparing the two models. One is the NN-AR model,

$$y_t^{(v)} = a_1^{(v)}y_{t-1}^{(v)} + \cdots + a_{p^{(v)}}^{(v)}y_{t-p^{(v)}}^{(v)} + \frac{b^{(v)}}{6}\sum_{v'\in N(v)} y_{t-1}^{(v')} + n_t^{(v)}$$
$$x_t^{(v)} = L^{-1}y_t^{(v)}$$

and another is the NN-ARX model,

$$y_t^{(v)} = a_1^{(v)}y_{t-1}^{(v)} + \cdots + a_{p^{(v)}}^{(v)}y_{t-p^{(v)}}^{(v)} + \frac{b^{(v)}}{6}\sum_{v'\in N(v)} y_{t-1}^{(v')} + \theta_1^{(v)}s_{t-1} + n_t^{(v)}$$
$$x_t^{(v)} = L^{-1}y_t^{(v)}$$

The significance of the extra term $\theta_1^{(v)} s_{t-1}$ can be checked by comparing the AIC or by checking the log-likelihood ratio of the two models. If we map the difference of the AIC of the two models at each voxel, we can see that the positive area of AIC, especially strongly positive area are showing the strongly activated area. An AIC plot shows, however, only the strength of the significance of the stimulus term, while if we plot the value of $\theta_1^{(v)} s_{t-1}$ at each voxel it shows "how" the stimulus is affecting the voxel, increasing the BOLD signal or reducing it.

5.3. Statistical check of instantaneous connectivities between remote voxels

Since the fMRI data are measured at a rather slow sampling rate, causal information transferred by neural connections may appear to be instantaneously driving the voxels from outside the two voxels. In other words, the prediction errors of the two remote voxels must be strongly correlated. The statistical significance of the simultaneous correlation of the prediction errors can be checked either by comparing AICs or checking the likelihood ratio of the following two models:

$$y_t^{(v)} = a_1^{(v)} y_{t-1}^{(v)} + \cdots + a_{p^{(v)}}^{(v)} y_{t-p^{(v)}}^{(v)} + \frac{b^{(v)}}{6} \sum_{v' \in N(v)} y_{t-1}^{(v')} + \theta_1^{(v)} s_{t-1} + n_t^{(v)}$$

$$y_t^{(w)} = a_1^{(w)} y_{t-1}^{(w)} + \cdots + a_{p^{(w)}}^{(w)} y_{t-p^{(w)}}^{(w)} + \frac{b^{(w)}}{6} \sum_{w' \in N(w)} y_{t-1}^{(w'')} + \theta_1^{(w)} s_{t-1} + n_t^{(w)}$$

$$x_t^{(v)} = L^{-1} y_t^{(v)}$$

$$x_t^{(w)} = L^{-1} y_t^{(w)} \quad \Sigma_n = \begin{pmatrix} \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \sigma_{vv} & \cdots & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & \cdots & \sigma_{ww} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

versus

$$y_t^{(v)} = a_1^{(v)} y_{t-1}^{(v)} + \cdots + a_{p^{(v)}}^{(v)} y_{t-p^{(v)}}^{(v)} + \frac{b^{(v)}}{6} \sum_{v' \in N(v)} y_{t-1}^{(v')} + \theta_1^{(v)} s_{t-1} + n_t^{(v)}$$

$$y_t^{(w)} = a_1^{(w)} y_{t-1}^{(w)} + \cdots + a_{p^{(w)}}^{(w)} y_{t-p^{(w)}}^{(w)} + \frac{b^{(w)}}{6} \sum_{w' \in N(w)} y_{t-1}^{(w'')} + \theta_1^{(w)} s_{t-1} + n_t^{(w)}$$

$$x_t^{(v)} = L^{-1} y_t^{(v)}$$

$$x_t^{(w)} = L^{-1} y_t^{(w)} \quad \Sigma_n = \begin{pmatrix} \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \sigma_{vv} & \cdots & \sigma_{vw} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \sigma_{vw} & \cdots & \sigma_{ww} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

Here, the innovations are given in the same way as for a standard NN-ARX model, but

$$\sum_t \left\{ \log \sigma^2_{\varepsilon_t^{(v)}} + \frac{(\varepsilon_t^{(v)})^2}{\sigma^2_{\varepsilon_t^{(v)}}} \right\} + \sum_t \left\{ \log \sigma^2_{\varepsilon_t^{(w)}} + \frac{(\varepsilon_t^{(w)})^2}{\sigma^2_{\varepsilon_t^{(w)}}} \right\}$$

in the (-2)log-likelihood (7) must be replaced by

$$\sum_t \left\{ \log \det \begin{pmatrix} \sigma^2_{\varepsilon_t^{(v)}} & \sigma_{vw} \\ \sigma_{vw} & \sigma^2_{\varepsilon_t^{(w)}} \end{pmatrix} + \begin{pmatrix} \varepsilon_t^{(v)} & \varepsilon_t^{(w)} \end{pmatrix} \begin{pmatrix} \sigma^2_{\varepsilon_t^{(v)}} & \sigma_{vw} \\ \sigma_{vw} & \sigma^2_{\varepsilon_t^{(w)}} \end{pmatrix}^{-1} \begin{pmatrix} \varepsilon_t^{(v)} \\ \varepsilon_t^{(w)} \end{pmatrix} \right\}$$

where

$$\sigma_{vw} = \frac{1}{N} \sum \varepsilon_t^{(v)} \varepsilon_t^{(w)}.$$

5.4. Statistical check of the dynamic correlations between remote voxels

Even though the hemodynamics between remote voxels have no physical interaction, they may appear to be correlated with a time lag, if the deoxygenation in voxel w systematically follows a few seconds after the deoxygenation in voxel v under a certain task condition. In such situations, the statistical significance of the dynamic correlations between the two voxels may be checked by comparing the AIC of the following two models or by checking the likelihood ratio of the two models.

$$y_t^{(v)} = a_1^{(v)} y_{t-1}^{(v)} + \cdots + a_{p^{(v)}}^{(v)} y_{t-p^{(v)}}^{(v)} + \frac{b^{(v)}}{6} \sum_{v' \in N(v)} y_{t-1}^{(v')} + \theta_1^{(v)} s_{t-1} + n_t^{(v)}$$

$$y_t^{(w)} = a_1^{(w)} y_{t-1}^{(w)} + \cdots + a_{p^{(w)}}^{(w)} y_{t-p^{(w)}}^{(w)} + \frac{b^{(w)}}{6} \sum_{w' \in N(w)} y_{t-1}^{(w')} + \theta_1^{(w)} s_{t-1} + n_t^{(w)}$$

$$x_t^{(v)} = L^{-1} y_t^{(v)}$$

$$x_t^{(w)} = L^{-1} y_t^{(w)}$$

versus

$$y_t^{(v)} = a_1^{(v)} y_{t-1}^{(v)} + \cdots + a_{p^{(v)}}^{(v)} y_{t-p^{(v)}}^{(v)} + \frac{b^{(v)}}{6} \sum_{v' \in N(v)} y_{t-1}^{(v')} + \theta_1^{(v)} s_{t-1} + n_t^{(v)}$$

$$y_t^{(w)} = a_1^{(w)} y_{t-1}^{(w)} + \cdots + a_{p^{(w)}}^{(w)} y_{t-p^{(w)}}^{(w)} + \frac{b^{(w)}}{6} \sum_{w' \in N(w)} y_{t-1}^{(w')} + c^{(w,v)} y_{t-1}^{(v)} + \theta_1^{(w)} s_{t-1} + n_t^{(w)}$$

$$x_t^{(v)} = L^{-1} y_t^{(v)}$$

$$x_t^{(w)} = L^{-1} y_t^{(w)}$$

5.5. Gaussianity of innovations

The guideline principle, the "Innovation Approach" or "Prediction Error Approach," is useful in various stage of modeling the dynamics, either microscopic or macroscopic, whenever the dynamic phenomena is measured in time series data. We know intuitionally that it is better for the prediction error to be small, and this is one of the reasons why the least squares method is widely used in the statistical analysis of time series. Prediction errors tend to become white noise since any temporal correlations are useful for further reducing the prediction errors.

A very important point we should pay attention to is the fact, summarized in Theorem 1, that the prediction errors are not only "white" but also "Gaussian" whenever the finite dimensional process is a sample-continuous finite-variance Markov (not necessarily Gaussian) process (for infinite dimensional stochastic processes, a similar theorem is given by K. Ito (1984)). The choice of the least squares criterion or the Gaussian innovation-based maximum likelihood criterion in many settings is mathematically supported by this theorem.

THEOREM 1. *(Doob, 1953; Feller, 1966) For any sample-continuous finite-variance d-dimensional Markov process x_t, the prediction error $v_t = x_t - E[x_t|x_{t-1}, x_{t-2}, \ldots, x_N, \theta]$ converges to a Gaussian white noise for $\Delta t \to 0$.*

If the prediction errors do not appear to be Gaussian, we need to reconsider the dynamic model which we are currently using. Often people tend to look for an easy solution, that is, to use the same dynamic model with more generally distributed non-Gaussian noise model. Errors with nonzero mean density distributions, asymmetric density distributions, fat-tailed density distributions, or bimodal density distributions sound more general and suitable than Gaussian errors, but actually this kind of idea is nonsensical. The non-Gaussian characters of the prediction errors are simply showing the inappropriateness of the dynamic model we assumed for the data. We must pay attention to the mathematical fact that the prediction errors "cannot" be too general. We already saw that Theorem 1 implies that the prediction errors of Markov diffusion processes are Gaussian. When the Markov process is not a diffusion type, that is, its sample path has discontinuous jumps, the Levy-Ito Decomposition Theorem says that the prediction error process is decomposed into two mutually independent processes, that is, Gaussian white noise and the (compensated) compound Poisson process (see Levy (1954) and Sato (1999) for details).

The most sensible and easy solution in this awkward situation for the prediction error-based maximum likelihood approach may be to redesign the experiment and eliminate all the possibility of pulse like shot noise and collect a new data set. If shot noise is unavoidable in the experiment or if the shot noise has some important physiological meaning, an alternative and natural extension of the prediction error-based maximum log-likelihood approach may be to use the Markov diffusion model with jumps, where the driving white Gaussian noise of the stochastic dynamical system is replaced by a sum of Gaussian white noise and a compound Poisson noise process. An approximate numerical solution for the maximum likelihood method for the detection of jumps

and estimation of models can be obtained (see Ozaki and Iino (2001), Jimenez and Carbonell (2006)), although the evaluation of the exact log-likelihood function for the Markov jump diffusion process with an observation error is quite difficult.

6. Applications to connectivity study and brain mapping

We note that, since the fMRI data are measured at a rather slow sampling rate, fast causal information transferred by neural connections may appear to be instantaneously driving the voxels from outside. Then, the NN-ARX model prediction errors of the two remote voxels, if they are systematically connected by a neural network, must have a strong correlation. A computational method to search for the pairs of the voxels, whose prediction errors have significantly large correlations between the pairs, leading to the further significant reduction of the (-2)log-likelihood, is already implemented in a toolbox, and applied to the analysis of fMRI data in some physiological experiments (see Bosch-Bayard et al. (2007, 2010)).

In brain data analysis, it always helps if the results are plotted on the brain image, and there are always two ways of plotting statistical results; one is the significance plot and another is the plot of "how" variables affect each other. For example, if we plot the difference of AIC of NN-AR model without an exogenous stimulus input variable and NN-ARX model with an exogenous stimulus input variable at each voxel, the map shows the strength of the significance of the activation at each voxel. However, the significance plot does not contain the information how the stimulus affects the BOLD signal at each voxel. If we plot the stimulus term $\theta_1^{(v)} s_{t-1}$ at each voxel, this could be more useful, since this will be showing whether it is contributing to increase the BOLD signal or decrease the BOLD signal. This may provide us with a useful phenomenological information for the understanding of the whole brain's responsive mechanism to the stimulus.

The same thing could be said for the plotting in the connectivity study. To see the connectivity between important voxels, say the primary visual cortex *V1* and the rest of the brain, we could compare two models: an NN-ARX model-1 which ignores simultaneous noise correlations between *V1* and the other voxels and an NN-ARX model-2 which takes account the simultaneous noise correlations between *V1* and the other voxels. We could either plot the difference of AIC of the two models, NN-ARX model-1 and NN-ARX model-2, at each voxel or plot the correlations between the prediction errors at *V1* and the prediction errors at the other voxel. The AIC plot shows the strength of significance of the simultaneous correlations of each voxel toward *V1*. However, it does not show whether the correlation is positive or negative, and it is always useful in understanding the brain function if we plot the correlation values at each voxel at the same time. The two plots, plot of the significance (difference of AICs) and the plot of the correlation values, compensate each other and they are to be used together by the analysts.

Incidentally, we note that the difference of the AIC of the two local models, dependent model and independent model for the two voxels v and w, is essentially equivalent

to the mutual information $I(x^{(v)}, x^{(w)}|\hat{\theta})$,

$$I\left(x^{(v)}, x^{(w)}|\hat{\theta}\right) = \log\left\{p\left(x_1^{(v)}, x_1^{(w)}\right)', \left(x_2^{(v)}, x_2^{(w)}\right)', \ldots \left(x_N^{(v)}, x_N^{(w)}\right)'|\hat{\theta}^{(vw)}\right\} - \log\left\{p\left(x_1^{(v)}, \ldots, x_N^{(v)}|\hat{\theta}^{(v)}\right) p\left(x_1^{(w)}, \ldots, x_N^{(w)}|\hat{\theta}^{(w)}\right)\right\}$$

Since we have

$$\begin{aligned} I\left(x^{(v)}, x^{(w)}|\hat{\theta}\right) &= \log\left\{p\left(x_1^{(v)}, x_1^{(w)}\right)', \left(x_2^{(v)}, x_2^{(w)}\right)', \ldots \left(x_N^{(v)}, x_N^{(w)}\right)'|\hat{\theta}^{(vw)}\right\} \\ &\quad - \log\left\{p\left(x_1^{(v)}, \ldots, x_N^{(v)}|\hat{\theta}^{(v)}\right) p\left(x_1^{(w)}, \ldots, x_N^{(w)}|\hat{\theta}^{(w)}\right)\right\} \\ &= \sum_{i=1}^{N} \log\left\{\frac{p\left(\left(\varepsilon_t^{(v|v,w)}, \varepsilon_t^{(w|v,w)}\right)'|\hat{\theta}^{(vw)}\right)}{p\left(\varepsilon_i^{(v)}|\theta^{(v)}\right) p\left(\varepsilon_i^{(w)}|\hat{\theta}^{(w)}\right)}\right\} \\ &= \left(-\frac{1}{2}\right) N\left[\log\left(1 - \hat{\rho}_{v,w}^2\right) + \left\{\left(\log\hat{\sigma}_{(v|v,w)}^2 - \log\hat{\sigma}_v^2\right)\right. \right. \\ &\quad \left.\left. + \left(\log\hat{\sigma}_{(w|v,w)}^2 - \log\hat{\sigma}_w^2\right)\right\}\right] + \text{Const} \end{aligned}$$

the plot of AIC difference or the plot of log of the likelihood ratio of the two models is essentially equivalent to the plot of $\log(1 - \hat{\rho}_{v,w}^2)$ of the prediction errors for v and w, where the sign (positive or negative) of $\hat{\rho}_{v,w}$ is lost.

Another merit of the innovation approach is related to the resolution of the information in the mapping. Even though the map of correlations between the whitened prediction errors $\varepsilon_1^{(v)}, \varepsilon_2^{(v)}, \ldots, \varepsilon_N^{(v)}$ and $\varepsilon_1^{(w)}, \varepsilon_2^{(w)}, \ldots, \varepsilon_N^{(w)}$ and between the original remote voxels $y_1^{(v)}, y_2^{(v)}, \ldots, y_N^{(v)}$ and $y_1^{(w)}, y_2^{(w)}, \ldots, y_N^{(w)}$ contain the same information of the instantaneous correlations, the resolution has been shown to be much clearer in the map of the correlations between the prediction errors than in the map of correlations between the original data in simulation studies (see Galka et al. (2006)). This means that the connectivity information, that is, significantly strong instantaneous correlations between the two remote connected voxels, can be elucidated and more easily seen in the prediction errors than may be seen in the original fMRI spatial time series.

7. Concluding remarks

In time series analysis, a natural way of characterizing the temporal correlation structure of a stationary time series is to use linear models, such as an AR model. For example, if the whole brain is in a stationary state, the multivariate 147,000 dimensional AR model of the fMRI (BOLD) signal determines the 147,000 dimensional autocovariance function, where we can extract autocorrelations at each voxel and cross

correlations between all the pair voxels. However, when the process is driven by an exogenous variable, the definition of the correlation structure of the process is not as straight forward as the stationary time series case.

The innovation approach is useful for finding any evidence of the need of more sophisticated modeling than simple stationary AR models. For example, the innovation approach is useful for detecting the existence of any effect of the stimulus variable by treating the stimulus indicator as an exogenous variable. By separating the effect of the stimulus from the ordinary behavior of the BOLD signals, we can estimate the correlation structure of the whole brain better than by the NN-AR model where the effect of the stimulus is interpreted as a part of the driving noise.

The validity of the innovation approach is not confined to linear modeling. If we notice any evidence from the prediction errors of the NN-ARX model at stimulus-on mode and prediction errors at the off mode, we could further generalize the NN-ARX model to a bilinear NN-ARX model such as

$$y_t^{(v)} = \sum_{i=1}^{p^{(v)}} \left(a_i^{(v)} + \alpha_i^{(v)} s_{t-1} \right) y_{t-i}^{(v)} + \frac{b^{(v)}}{6} \sum_{v' \in N(v)} y_{t-1}^{(v')} + \theta_1^{(v)} s_{t-1} + n_t^{(v)}$$

$$x_t^{(v)} = L^{-1} y_t^{(v)}.$$

AIC and log-likelihood ratio statistics are always useful for checking the significance of the contribution of the extra terms in the generalized model.

Once we find several important pairs of remote voxels having strong correlations through the above mentioned exploratory brain mapping methods, we can do further more elaborate analysis of dynamic causality between the remote voxels using techniques developed in time series analysis (Akaike, 1968, 1974; Granger, 1969; Yamashita et al., 2005). Especially, Akaike's parametric spectral method (Akaike, 1968; Ozaki, 2012) is useful in finding the causality from one variable to another through unobserved hidden state variables in complicated high-dimensional feedback systems. Here, the state space representation approach becomes useful for the situation where the driving noise is strongly correlated between the variables in the feedback system (Ozaki, 2012; Wong and Ozaki, 2006).

Acknowledgement

The ideas in this chapter have been developed in discussion with Dr. J. Bosch-Bayard, Prof. R. Biscay, Dr. A. Galka, Dr. K.F.K. Wong, Dr. J. Riera, Prof. R. Kawashima, Prof. N. Sadato, and Prof. P. Valdes-Sosa. The author is especially grateful to Prof. N. Sadato of the National Institute of Physiological Sciences and Dr. J. Bosch-Bayard of Cuban Neuroscience Center for their stimulating discussions and support.

References

Akaike, H., 1968. On the use of a linear model for the identification of feedback systems. Ann. Inst. Statist. Math. 20, 425–439.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contri. AC-19, 6, 7116–723.

Bosch-Bayard, J., Riera-Diaz, J.J., Biscay-Lirio, R., Wong, K.F.K., Galka, A., Yamashita, O., et al., 2007. Spatio-temporal correlations in fMRI time series: the whitening approach. Research Memo No.1025, Institute of Statistical Mathematics, Tokyo.

Bosch-Bayard, J., Riera-Diaz, J.J., Biscay-Lirio, R., Wong, K.F.K., Galka, A., Yamashita, O., et al., 2010. Spatio-temporal correlations from fMRI time series based on the NN-ARx Model. J. Integr. Neurosci. 9, 381–406.

Box, G.E.P., Jenkins, 1970. Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco.

Brown, P.E., Karesen, K.F., Roberts G.O., Tonellato, S., 2000. Blur-generated non-separable space-time models. J.R.Statist. Soc. B. 62(4), 847–860.

Buxton, R.B., 2002. An Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques Cambridge University Press, Cambridge.

Cressie, N., 1993. Statistics for Spatial Data. Wiley, New York.

Doob, J.L., 1953. Stochastic Processes. John Wiley & Sons, New York.

Feller, W., 1966. An Introduction to Probability Theory and Its Applications. John Wiley & Sons, New York.

Friston, K.J., Buckel, C., 2004. Functional connectivity. In: Frackowiak, R.S.J., Friston, K.J., Frith, C., Dolan, R., Price, C.J., Zeki, S., Ashburner, J., Penny, W.D. (Eds.), Human Brain Function, second ed. Academic Press, San Diego, pp. 999–1018.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., Frackowiak, R.S.J., 1995. Statistical parametric maps in functional imaging: A general linear approach. Hum. Brain Mapp. 2, 189–210.

Galka, A., Ozaki, T., Bosch-Bayard, J., Yamashita, O., 2006. Whitening as a tool for estimating mutual information in spatiotemporal datasets. J. Stat. Phys. 124(5), 1275–1315.

Galka, A., Yamashita, O., Ozaki, T., Biscay, R., Valdes-Sosa, P., 2004. A solution to the dynamical inverse problem of EEG generation using spatiotemporal Kalman filtering. NeuroImage 23, 435–453.

Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37(3), 424–438.

Ito, K., 1984. Infinite dimensional Ornstein-Uhlenbeck processes. In: Ito, K. (Ed.), Taniguchi Symposium SA, Katata, 1982, North-Holland, pp. 197–224.

Jimenez, J.C., Carbonell, F., 2006. Local linear approximations of jump diffusion processes. J. Appl. Prob. 43, 185–194.

Lange, N., Zeger, S.L., 1997. Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging (with discussion). Appl. Statist. 46, 1–29.

Levy, P., 1954. Theorie de l'Addition des Variables Aleatoires, second ed. Gauthier-Villars, Paris.

Ozaki, T., 1998. Dynamic X-11 model and nonlinear seasonal adjustment—II: numerical examples and discussion. Proc. Institut of Statist. Math., 45, 287–300 (in Japanese).

Ozaki, T., 2012. Time Series Modelling of Neuroscience Data. Chapman Hall, London.

Ozaki, T., Iino, M., 2001. An innovation approach to non-Gaussian time series analysis. J. App. Prob. Trust 38A, 78–92.

Riera, J., Bosch, J., Yamashita, O., Kawashima, R., Sadato, N., Okada, T., et al., 2004. fMRI activation maps based on the NN-ARx model. Neuroimage 23, 680–697.

Sato, K., 1999. Levy Processes and Infinitely Divisible Distributions. Cambridge University Press, Cambridge.

Whittle, P., 1962. Topographic correlation, power-law covariance function, and diffusion. Biometrika 49, 305–314.

Wiener, N., 1949. Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications. Wiley, New York (Originally issued as a classified report by MIT Radiation Lab., Cambridge, February 1942).

Wong, K.F.K., Ozaki, T., 2006. Akaike causality in state space—Instantaneous causality between visual cortex in fMRI time series.Biol. Cybern. 97, 151–157.

Worsley, K.J., Liao, C.H., Aston, J., Petre, V., Duncan, G.H., Morales, F., et al., 2002. A general statistical analysis for fMRI data. Neuroimage 15, 1–15.

Yamashita, O., Galka, A., Ozaki, T., Biscay, R., Valdes-Sosa, P., 2004. Recursive penalized least squares solution for dynamical inverse problems of EEG generation. Hum. Brain Mapp. 21, 221–235.

Yamashita, O., Sadato, N., Okada, T., Ozaki, T., 2005. Evaluating frequency-wise directed connectivity of BOLD signals applying relative power contribution with the linear multivariate time series models. Neuroimage 25, 478–490.

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00012-0

Count Time Series Models

Konstantinos Fokianos
Department of Mathematics & Statistics, University of Cyprus, Nicosia 1678, Cyprus

Abstract

We review regression models for count time series. We discuss the approach that is based on generalized linear models and the class of integer autoregressive processes. The generalized linear models' framework provides convenient tools for implementing model fitting and prediction using standard software. Furthermore, this approach provides a natural extension to the traditional ARMA methodology. Several models have been developed along these lines, but conditions for stationarity and valid asymptotic inference were given in the literature only recently. We review several of these facts. In addition, we consider integer autoregressive models for count time series and discuss estimation and possible extensions based on real data applications.

Keywords: autocorrelation, link function, Poisson distribution, prediction, stationarity.

1. Introduction

Figure 1 motivates the study of appropriate models for the statistical analysis of count time series. The upper left plot shows a time series of claims – referred to as series C3 – of short-term disability benefits made by cut-injured workers in the logging industry (Zhu and Joe, 2006). The lower left plot shows the usual sample autocorrelation function (ACF) for these data. If the ACF is adapted as a measure of correlation between pairs of observations, then the resulting plot points to weak correlation that decays fast, after a few lags. Therefore, to model these data, in the spirit of usual ARMA models (see Brockwell and Davis (1991) for instance), a few lagged variables entertained by a regression model will suffice to describe the correlation among the data. The right plots illustrate quite the opposite situation. The upper right plot of Fig. 1 shows the

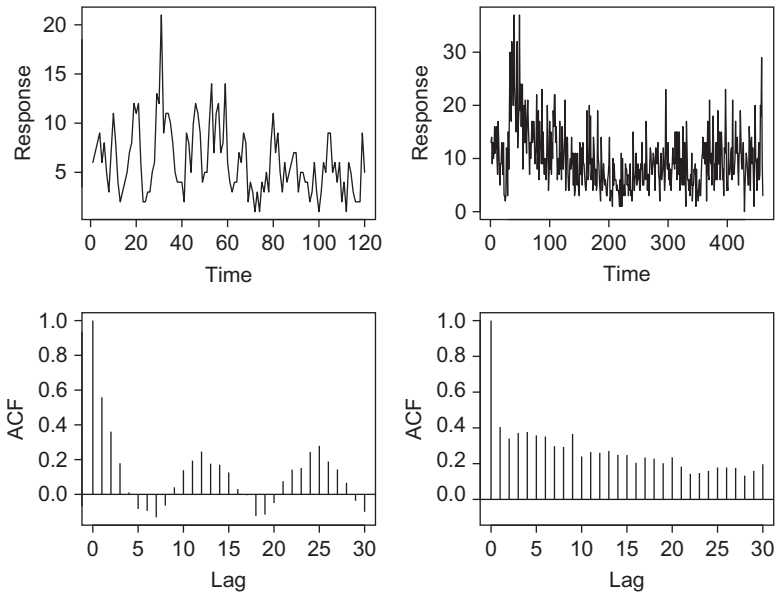

Fig. 1. Left plot: Time series of claims (up) and their sample ACF (bottom). Right plot: Number of transactions (up) and their sample ACF (bottom).

number of transactions for the stock Ericsson B, for one-day period. The lower right plot of the same figure shows the sample ACF for these data. When compared to the previous data example, we note a distinct feature characterizing the transactions data; there exists a strong correlation among observations, which decays slowly. Hence, a few lagged variables in a regression model will not be sufficient to accommodate these particular features of the data. In other words, the modeling of these data raises analogous questions and poses similar challenges to the case of ARCH and GARCH models (Bollerslev, 1986; Engle, 1982). The main goal of this chapter is to discuss statistical inference for count time series and give some guidelines for inference in situations similar to the aforementioned data examples. By doing so, we also review some important probabilistic properties of such models and the associated statistical inference.

Modeling counts of events can be found in all areas of statistics and econometrics, and throughout the social and physical sciences. Apart from the above cases, we can observe daily number of hospital admissions, monthly number of cases of some disease, weekly number of rainy days, and so on. For the regression analysis of count data, the ordinary linear model would not be applicable, because the response variable assumes discrete values. However, a related counterpart is the Poisson regression model and it is a natural starting point to extend it to dependent count data. The Poisson regression model has been used in several applied areas to model counts. In fact, the Poisson model is a nonlinear, albeit straightforward, and popular modeling tool, whose fitting is implemented by standard software. This chapter will survey models and methods for analyzing time series of counts, beginning with this basic tool.

The Poisson model provides the main instrument for modeling count time series data. However, other distributional assumptions may be used instead; the most natural among other candidates being the negative binomial distribution. Regardless of the chosen distribution, we will review mostly models that fall under the framework of generalized linear models for time series. This class of models and the maximum likelihood theory provide a systematic framework for the analysis of quantitative as well as qualitative time series data. Indeed, estimation, diagnostics, model assessment, and forecasting are implemented in a straightforward manner, where the computation is carried out by a number of existing software packages. Experience with these models shows that both positive and negative association can be taken into account by a suitable parametrization of the model. These issues are addressed in the list of desiderata suggested by Davis et al. (1999) and Zeger and Qaqish (1988).

There are other alternative classes of regression models for count time series; the most prominent being the integer autoregressive models. These models are based on the notion of thinning operator. Accordingly, integer autoregressive models imitate the structure of the common autoregressive process in the sense that the thinning operation is applied instead of scalar multiplication.

This chapter surveys several of the above models. We discuss their properties, estimation methods, and theory. Section 2 is introductory to Poisson regression modeling. Section 3 discusses, in detail, linear and log-linear models for count time series. The Poisson assumption is dropped in Section 4, where we study models using different distributional assumptions. Section 5 summarizes properties of the integer autoregressive models. Finally, Section 6 concludes this work with other potential applications and further development of the methodology. For ease of presentation, we use slightly abused notation for the regression parameters and the error sequences. However, this does not affect the main concepts as will be clear from the context.

2. Poisson regression modeling

The Poisson distribution is commonly used to model rate of random events that occur (arrive) in some fixed time interval. If we assume that λ denotes the rate of arrivals, then the distribution of the random variable Y, which denotes the number of arrivals in a fixed time interval, follows the Poisson distribution with probability mass function

$$\mathrm{P}[Y = y] = \frac{\exp(-\lambda)\lambda^y}{y!}, \quad y = 0, 1, 2, \ldots \tag{1}$$

It is an elementary exercise to show that the mean and variance of Y are both equal to λ; $\mathrm{E}[Y] = \mathrm{Var}[Y] = \lambda$. In fact, this property characterizes the Poisson distribution. A related property is that the cumulant generating function of a Poisson random variable is given by $K_Y(t) \equiv \log M_Y(t) = \lambda(\exp(t) - 1)$, where $M_Y(t)$ is the moment generating function of Y. This can be proved by simple calculations, but for this presentation, it is instructive to consider the Poisson distribution as a member of the natural exponential family of distributions.

Let $f(x;\theta)$ denote the density function of the natural exponential family with parameter θ, i.e.,

$$f(x;\theta) = h(x)\exp(\theta x - b(\theta)), \quad x \in \mathcal{A}, \tag{2}$$

where $h(\cdot)$, $b(\cdot)$ are known functions and $\mathcal{A}$ is a subset of $\mathbb{R}$. Then, it is straightforward to show that the Poisson distribution is expressed as in (2) with $\theta = \log\lambda$, $b(\theta) = \exp(\theta)$, and $h(x) = 1/x!$. Using the fact that the cumulant generating function of (2) is equal to $b(t+\theta) - b(\theta)$, the claim follows.

In most of the applications, count data are usually observed with some covariate information. For example, see the works of McCullagh and Nelder (1989, Section 6.3.2) where the authors study the relation between the type of ship, its year of construction, and its service period to the expected number of damage incidents using the logarithm of the aggregate months of service as an offset. (An offset is a continuous regression variable with corresponding known regression coefficient equal to 1.) In general, assume that $X_1, \ldots, X_p$ are p regression variables observed jointly with a count response variable Y that follows the Poisson distribution. A possible regression model for association between the regressors and the expected value of Y given $X_1, \ldots, X_p$ is

$$\lambda = \beta_0 + \sum_{i=1}^{p} \beta_i X_i. \tag{3}$$

This is an ordinary linear model with unknown regression coefficients β_i, $i = 0, \ldots, p$ to be estimated. Model (3) poses several difficulties for fitting, because the parameter λ has to be positive. Nevertheless, in the context of time series and when the correlation among successive observations is positive, models such as (3) are quite useful; recall Fig. 1. A more natural choice for the regression modeling of count data is the so called log-linear model which is specified by

$$\log\lambda = \beta_0 + \sum_{i=1}^{p} \beta_i X_i, \tag{4}$$

where the notation is as in (3). Regardless of the chosen model, a fact that remains true is that both (3) and (4) belong to the class of generalized linear models as introduced by Nelder and Wedderburn (1972) and elaborated further by McCullagh and Nelder (1989). Recall, that a generalized linear model consists of three components; the random component that belongs to the exponential family of distributions (2) with $\mathrm{E}[X] = \mu$, the systematic component η, and the link function $g(\cdot)$. The link function is a monotone twice differentiable function that is chosen by the user (or can be estimated). This function associates the random and systematic component via $g(\mu) = \eta$. For the Poisson distribution, it is clear that both (3) and (4) introduce a generalized linear model with $\eta = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$ and $g(\lambda) = \lambda$ (for (3)) and $g(\lambda) = \log\lambda$ (for (4)). Estimation and inference are based on the maximum likelihood theory – this topic has been described in several texts; see McCullagh and Nelder (1989) and Agresti (2002), for example. In the next section, we explore these ideas in the context of count time series.

3. Poisson regression models for count time series

It is useful to consider the classical AR(1) process

$$Y_t = b_1 Y_{t-1} + \epsilon_t, \tag{5}$$

where $|b_1| < 1$ and $\{\epsilon_t\}$ is a sequence of independent and identically distributed (i.i.d.) normal random variables with zero mean and variance σ^2. This is a standard model used for the analysis of real-valued time series. It implies that the value of the process at time t depends on the value of the process at time $(t-1)$ plus a random error; e.g., Priestley (1981), Brockwell and Davis (1991, Chapter 3), and Shumway and Stoffer (2006). It is enlightening to consider model (5) as a member of the family of generalized linear models for time series. Recalling the discussion at the end of the last section, note that the random component of the model (for the AR(1) model (5) the conditional probability density function of Y_t given its past is Gaussian) belongs to the exponential family of distribution. In addition, the systematic component is defined by $\eta_t = b_1 Y_{t-1}$. If the link function $g(\cdot)$ is chosen to be the identity, then $g(\mathrm{E}[Y_t \mid Y_{t-1}] = \eta_t$. Hence, the AR(1) process (5) falls within the framework of generalized linear models for time series, see Kedem and Fokianos (2002, Chapter 1). This discussion motivates much of the following development.

3.1. Linear models for count time series

From this point on, assume that $\{Y_t\}$ denotes a count time series; we will call this process the "response." Following the AR(1) paradigm, we generalize model (5) in the context of Poisson autoregression by assuming that

$$Y_t \mid \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \quad \lambda_t = d + b_1 Y_{t-1}, \quad t \geq 1, \tag{6}$$

with $\mathcal{F}_t = \sigma(Y_s, s \leq t)$, d, b_1 non-negative parameters and $\{\lambda_t\}$ denoting the mean process of Y_t given its past. Positive d and b_1 ensure that $\lambda_t > 0$, since Y_t is a non-negative integer. With this notation, it is clear that model (6) falls within the framework of generalized linear models with the random component being the Poisson distribution, the systematic component given by $\eta_t = d + b_1 Y_{t-1}$ and the identity link. This is a situation quite analogous to (5). Model (6) implies the same dynamics of model (5), since

$$Y_t = \lambda_t + (Y_t - \lambda_t) = d + b_1 Y_{t-1} + \epsilon_t, \quad t \geq 1, \tag{7}$$

where the notation is obvious. The last line displays that the values of the process at time t depend on the value of the process at time $(t-1)$ plus the term $\{\epsilon_t\}$, which is white noise sequence; that is a sequence of uncorrelated random variables with zero mean and constant variance. Indeed, if we assume that the process $\{Y_t\}$ is stationary, then we obtain the following results:

- Constant mean:

$$\mathrm{E}[\epsilon_t] = \mathrm{E}\Big[(Y_t - \lambda_t)\Big] = \mathrm{E}\Big[\mathrm{E}\Big(Y_t - \lambda_t \mid \mathcal{F}_{t-1}\Big)\Big] = 0,$$

- Constant variance:

$$\mathrm{Var}[\epsilon_t] = \mathrm{Var}\Big[\mathrm{E}\Big(\epsilon_t \mid \mathcal{F}_{t-1}\Big)\Big] + \mathrm{E}\Big[\mathrm{Var}\Big(\epsilon_t \mid \mathcal{F}_{t-1}\Big)\Big] = \mathrm{E}[\lambda_t] = \mathrm{E}[Y_t].$$

  This is independent of t since $\{Y_t\}$ has been assumed to be stationary. The last equality follows from the fact that $\mathrm{E}[\epsilon_t] = 0$.
- Uncorrelated sequence: For $k > 0$,

$$\mathrm{Cov}(\epsilon_t, \epsilon_{t+k}) = \mathrm{E}[\epsilon_t \epsilon_{t+k}] = \mathrm{E}\Big[\epsilon_t \mathrm{E}\Big(\epsilon_{t+k} \mid \mathcal{F}_{t+k-1}\Big)\Big] = 0$$

These results verify the claim that the sequence $\{\epsilon_t\}$ is a white noise sequence. Because of the assumed stationarity, (7) shows that $\mathrm{E}[Y_t] = d + b_1 \mathrm{E}[Y_{t-1}]$ and therefore that $\mathrm{Var}[\epsilon_t] = \mathrm{E}[Y_t] = d/(1 - b_1)$; a fact which illustrates that b_1 needs to be positive but less than 1.

To start investigating the second-order properties of (6), we employ representation (7). By repeated substitution, we obtain that

$$\begin{aligned} Y_t &= d + b_1 Y_{t-1} + \epsilon_t \\ &= d + b_1(d + b_1 Y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\ &= d(1 + b_1) + b_1^2 Y_{t-2} + b_1 \epsilon_{t-1} + \epsilon_t \\ &= \cdots\cdots\cdots\cdots\cdots \\ &= d(1 + b_1 + b_1^2 + \cdots b_1^t) + \sum_{i=0}^{t} b_1^i \epsilon_{t-i}. \end{aligned} \tag{8}$$

Therefore, as in the case of the usual AR(1) model, assuming that $0 < b_1 < 1$, we obtain by (8) for large t, the useful representation

$$Y_t = \frac{d}{1 - b_1} + \sum_{i=0}^{\infty} b_1^i \epsilon_{t-i},$$

in mean square sense. Standard arguments now show that the autocovariance function of model (6) is given by

$$\mathrm{Cov}(Y_t, Y_{t+h}) = \frac{b_1^h}{1 - b_1^2} \mathrm{E}[Y_t], \quad h \geq 0,$$

a fact that yields the ACF of model (6):

$$\mathrm{Corr}(Y_t, Y_{t+h}) = b_1^h, \quad h \geq 0. \tag{9}$$

Note that unless $b_1 = 0$, the variance of $\{Y_t\}$ is always greater than its expectation; i.e., model (6) takes into account overdispersion. These results are straightforward consequences of (7) because it reveals that (6) has identical second-order properties to

those of the AR(1) model (5). However, $\text{Corr}(Y_t, Y_{t+h}) > 0$, for all $h > 0$, because $b_1 > 0$; that is, model (6) can be employed for positively correlated count time series.

An empirical verification of these considerations is illustrated in the left plot of Fig. 2. The upper plot shows 200 observations from (6) with $d = 1$ and $b_1 = 0.6$. The lower plot shows the autocorrelation function of the same model. Quite clearly, as the lag h increases, the autocorrelation function tends fast to smaller values; see Eq. (9) and compare this plot with the left-hand plot of Fig. 1. (Further results about moments and cumulants of model (6) are given in Weiß (2010)).

The right plot of Fig. 2 illustrates a different situation. It shows 200 realizations of the following model

$$Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda} \sim \text{Poisson}(\lambda_t), \quad \lambda_t = d + a_1\lambda_{t-1} + b_1 Y_{t-1}, \quad t \geq 1, \tag{10}$$

where $\mathcal{F}_t^{Y,\lambda}$ the σ-field generated by $\{Y_0, \ldots, Y_t, \lambda_0\}$, that is, $\mathcal{F}_t^{Y,\lambda} = \sigma(Y_s, \lambda_0, s \leq t)$, and $\{\lambda_t\}$ is a Poisson intensity process, as before. The parameters d, a_1, b_1 are assumed to be positive and to satisfy $0 < a_1 + b_1 < 1$. Both starting values λ_0 and Y_0 are assumed to be random. When $a_1 = 0$, then model (6) is recovered. Recall the lower right plot of Fig. 2 and compare it with the corresponding plot of Fig. 1. Apparently, the persistence of large positive values of the autocorrelation function is a consequence of the existence of the feedback mechanism $\{\lambda_t\}$ introduced in (10). In principle, when count time series are available and their autocorrelation function assumes relatively high values for large lags, then we should expect a model of the form (6) to accommodate this fact by entertaining a large number of lagged regressor variables. However,

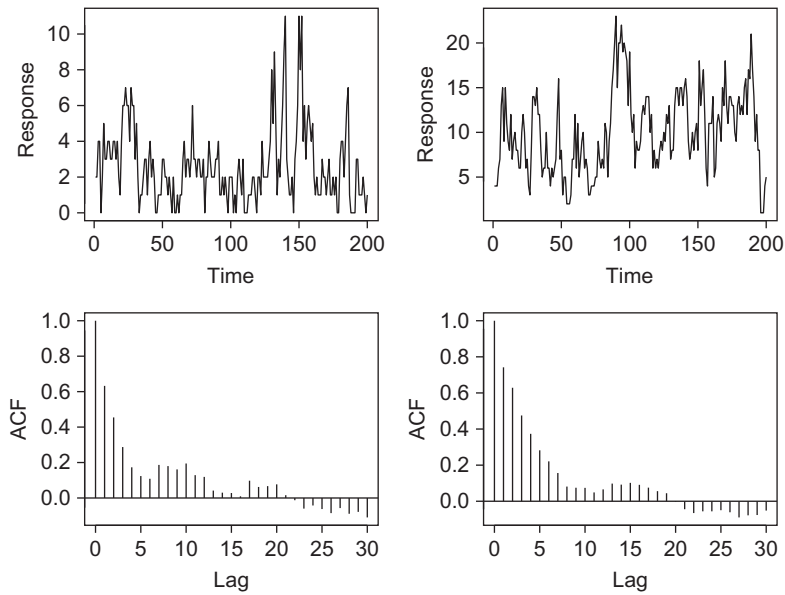

Fig. 2. Left plot: Two hundred observations (up) and their sample ACF (bottom) from model (6) for $d = 1$ and $b_1 = 0.6$. Right plot: Two hundred observations (up) and their sample ACF (bottom) from model (10) for $d = 1$, $a_1 = 0.3$, and $b_1 = 0.6$.

such an approach can be avoided when employing model (10); it simply provides a parsimonious way to model this type of data (Fokianos et al., 2009).

Several results for model (10) have been reported in the literature, see Rydberg and Shephard (2000), Streett (2000), Heinen (2003), and Ferland et al. (2006), who consider the following general model of order (p, q):

$$Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda} \sim \text{Poisson}(\lambda_t), \quad \lambda_t = d + \sum_{i=1}^{p} a_i \lambda_{t-i} + \sum_{j=1}^{q} b_j Y_{t-j}, t \geq \max(p, q), \tag{11}$$

and show that it is second-order stationary provided that $0 < \sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_j < 1$.

To study the properties of (10), it is instructive to consider again decomposition (7) and then use the second part of (10) to express the response process as

$$Y_t = d + (a_1 + b_1)Y_{t-1} + \epsilon_t - a_1\epsilon_{t-1},$$

with some slight abuse of notation. In the last display, $\epsilon_t = Y_t - \lambda_t$ and this sequence of random variables, although distinct from the corresponding sequence defined by means of (7), is still a white noise process; the proof of this fact is the same as to the case of the noise sequence that corresponds to model (6).

Furthermore, the last display can be rewritten as

$$\left(Y_t - \frac{d}{1 - (a_1 + b_1)}\right) = (a_1 + b_1)\left(Y_{t-1} - \frac{d}{1 - (a_1 + b_1)}\right) + \epsilon_t - a_1\epsilon_{t-1}, \tag{12}$$

which shows that (10) has exactly identical second-order properties as those of a usual ARMA(1,1) model. Hence, when $0 < a_1 + b_1 < 1$, then there exists a stationary solution $\{Y_t\}$ of (10), with mean $\mathrm{E}[Y_t] = \mathrm{E}[\lambda_t] \equiv \mu = d/(1 - a_1 - b_1)$ and autocovariance function

$$\mathrm{Cov}[Y_t, Y_{t+h}] = \begin{cases} \dfrac{(1 - (a_1 + b_1)^2 + b_1^2)\mu}{1 - (a_1 + b_1)^2}, & h = 0, \\ \dfrac{b_1(1 - a_1(a_1 + b_1))(a_1 + b_1)^{h-1}\mu}{1 - (a_1 + b_1)^2}, & h \geq 1. \end{cases}$$

It is clear that the ACF of model (10) is equal to

$$\mathrm{Corr}[Y_t, Y_{t+h}] = \frac{b_1(1 - a_1(a_1 + b_1))(a_1 + b_1)^{h-1}}{(1 - (a_1 + b_1)^2 + b_1^2)}, \quad h \geq 1.$$

This fact matches the right-hand plots of Figs. 1 and 2 and explains the slower decay of the corresponding ACF.

For the Poisson distribution, $\mathrm{E}[Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda}] = \mathrm{Var}[Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda}] = \lambda_t$. Therefore, model (10) can be defined as an INGARCH(1,1); i.e., an integer GARCH model, because its structure is analogous to that of the customary GARCH model whereby volatility is regressed on past values of itself and squared responses. In fact, model (11) can be termed as an INGARCH(p, q) model. However (10), and more generally (11), specify a *conditional mean relation* to the past values of both λ_t and Y_t. Observe that $\mathrm{Var}[Y_t] \geq \mathrm{E}[Y_t]$ with equality when $b_1 = 0$. Thus, the inclusion of the past values of Y_t in the evolution of λ_t yields overdispersion – this is the same fact that holds true for model (6). Furthermore, $\mathrm{Corr}\,[Y_t, Y_{t+h}] > 0$ for model (10) like in the case of model (6).

By repeated substitution,

$$\begin{aligned}
\lambda_t &= d + a_1\lambda_{t-1} + b_1 Y_{t-1} \\
&= d + a_1(d + a_1\lambda_{t-2} + b_1 Y_{t-2}) + b_1 Y_{t-1} \\
&= d + a_1 d + a_1^2\lambda_{t-2} + a_1 b_1 Y_{t-2} + b_1 Y_{t-1} \\
&= \cdots\cdots\cdots\cdots\cdots \\
&= d\frac{1 - a_1^t}{1 - a_1} + a_1^t\lambda_0 + b_1\sum_{i=0}^{t-1} a_1^i Y_{t-i-1}. \qquad (13)
\end{aligned}$$

The last display shows that the hidden process $\{\lambda_t\}$ is determined by past functions of lagged responses and the initial value λ_0. Therefore, model (10) belongs to the class of observation driven models in the sense of Cox (1981). Representation (13) explains further the reason that model (10) offers a parsimonious way of modeling count time series data whose ACF decays slowly; see the example shown in the right plot of Fig. 1. The process $\{\lambda_t\}$ depends on a large number of lagged response values, so it is expected to provide a more parsimonious model than a model of the form (6).

As a final remark, when $a_1 + b_1$ approaches 1, then the ACF function of model (10) becomes unstable and the resulting model has similar properties to those of an integrated GARCH model; that is, predictions for λ_t will reflect the most recent variation found in the data. Such models have not been studied in the literature.

3.2. *Log-linear models for count time series*

The previous discussion shows that model (10) provides a satisfactory conceptual framework for modeling-dependent count data. However, the model definition imposes implicitly some restrictions on the data. First recall that, $\mathrm{Cov}\,[Y_t, Y_{t+h}] > 0$, because $0 < a_1 + b_1 < 1$. Therefore, model (10) cannot be employed for modeling negative correlation among successive observations. An additional drawback of (10) is that it does not accommodate covariates in a straightforward way, because of the identity link function. However, as it was mentioned in Section 2, the choice of the logarithmic function is the most popular among the link functions for modeling count data. In fact, this choice corresponds to the canonical link model. Hence, we resort to log-linear models for count time series; see Zeger and Qaqish (1988), Li (1994), MacDonald and Zucchini (1997), Brumback et al. (2000), Kedem and Fokianos (2002), Benjamin et al. (2003), Davis et al. (2003), Fokianos and Kedem (2004), Jung et al. (2006), Creal et al. (2008), and Fokianos and Tjøstheim (2011).

Suppose again that $\{Y_t\}$ denotes a count time series. We will be working with the so-called canonical link process $\nu_t \equiv \log \lambda_t$. We study the following family of log-linear autoregressive models

$$Y_t \mid \mathcal{F}_{t-1}^{Y,\nu} \sim \text{Poisson}(\lambda_t), \quad \nu_t = d + a_1 \nu_{t-1} + b_1 \log(Y_{t-1} + 1), \quad t \geq 1. \tag{14}$$

where $\mathcal{F}_t^{Y,\nu}$ the σ-field generated by $\{Y_0, \ldots, Y_t, \nu_0\}$, that is, $\mathcal{F}_t^{Y,\nu} = \sigma(Y_s, \nu_0, s \leq t)$. In general, the parameters d, a_1, b_1 can be positive or negative but they need to satisfy certain conditions so that we obtain a stationary time series. Both ν_0 and Y_0 are assumed again to be some random starting values.

Note that the lagged observations of the response Y_t are fed into the autoregressive equation for ν_t via the term $\log(Y_{t-1} + 1)$. This is a one-to-one transformation of Y_{t-1}, which is quite standard in coping with zero data values. Moreover, both λ_t and Y_t are transformed into the same scale. Covariates can be accommodated by model (14), by including them in the second equation of (14). An alternative modeling approach is based upon employing the transformation $\log(\max(Y_{t-1}, c))$, (cf. Zeger and Qaqish, 1988) for $c \in (0, 1]$, instead of $\log(Y_{t-1} + 1)$ in (14).

When $a_1 = 0$, we obtain the model

$$\nu_t = d + b_1 \log(Y_{t-1} + 1), \quad t \geq 1, \tag{15}$$

which parallels the structure of (6). With this notation, it is clear that model (15) falls within the framework of generalized linear models with the random component being the Poisson distribution, the systematic component given by $\eta_t = d + b_1 \log(Y_{t-1} + 1)$ and the link function being the logarithmic. Figure 3 illustrates the same phenomenon as that observed in Fig. 2, namely the inclusion of the feedback mechanism yields parsimony when the correlation decays slowly to zero.

The log-intensity process of (14) can be rewritten as

$$\nu_t = d \frac{1 - a_1^t}{1 - a_1} + a_1^t \nu_0 + b_1 \sum_{i=0}^{t-1} a_1^i \log(1 + Y_{t-i-1}), \tag{16}$$

after repeated substitution. Hence, we obtain again that the hidden process $\{\nu_t\}$ is determined by past functions of lagged responses. Equivalently, the log-linear model (14) belongs to the class of observation driven models and possess similar properties to the linear model (10).

To motivate further the choice of the $\log(\cdot)$ function for the lagged values of the response, consider a model like (14), but with Y_{t-1} included instead of $\log(Y_{t-1} + 1)$. In other words, set

$$Y_t \mid \mathcal{F}_{t-1}^{Y,\nu} \sim \text{Poisson}(\lambda_t), \quad \nu_t = d + a_1 \nu_{t-1} + b_1 Y_{t-1}.$$

In this case

$$\lambda_t = \exp(d) \lambda_{t-1}^{a_1} \exp(b_1 Y_{t-1}),$$

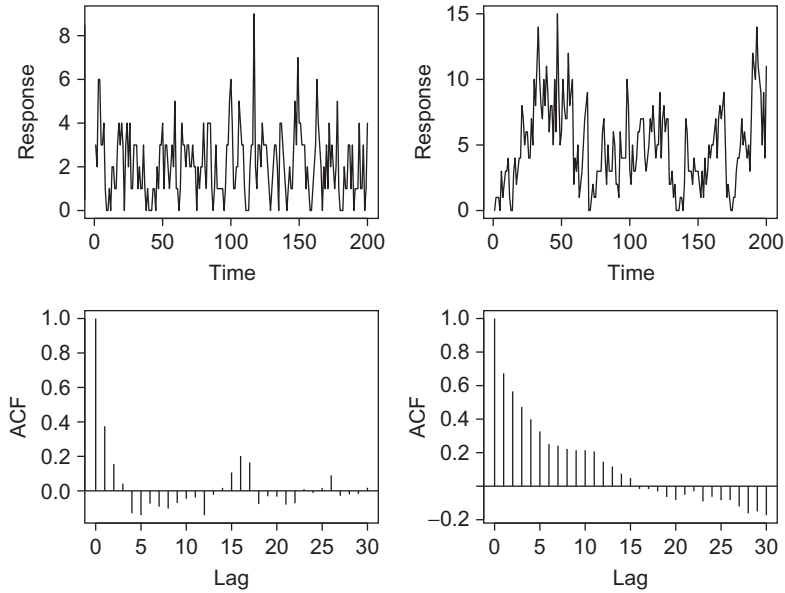

Fig. 3. Left plot: Two hundred observations (up) and their sample ACF (bottom) from model (15) for $d = 0.1$ and $b_1 = 0.6$. Right plot: Two hundred observations (up) and their sample ACF (bottom) from model (14) for $d = 0.1$, $a_1 = 0.3$, and $b_1 = 0.6$.

and therefore stability of the above system is guaranteed only when $b_1 < 0$. Otherwise, the process $\{\lambda_t\}$ increases exponentially fast, see Wong (1986) and Kedem and Fokianos (2002, Chapter 4) for more details. Hence, only negative correlation can be introduced by such a model. However (14) yields both positive (respectively, negative) correlation by allowing the parameter b_1 to take positive (respectively, negative) values. It is a challenging problem to obtain an explicit expression for the autocorrelation function of model (14). This is easily seen by considering (16). Exponentiating both sides of this formula shows that

$$\lambda_t = \exp\Big(d(1 - a_1^t)/(1 - a_1)\Big)\lambda_0^{a_1^t} \prod_{i=0}^{t-1} \Big(1 + Y_{t-i+1}\Big)^{b_1 a_1^i},$$

which demonstrates the complications of calculating first and second moments for model (14). However, by simulating a very long path of the series, we get a clue of the range of possible values of correlation obtained by (14). Table 1 illustrates the ACF of model (14) at lags one and two. It is evident that the log-linear model takes into account both negative and positive correlations.

An alternative log-linear model specification for count time series was studied by Davis et al. (2003). The model is given by the following

$$\nu_t = \beta_0 + \sum_{i=1}^{p} \beta_i \zeta_{t-i}, \tag{17}$$

Table 1
Typical values of the autocorrelation function at lags 1 and 2 derived by model (14) for selected values of the parameters a_1 and b_1 when $d = 0.5$. Results are based on 10,000 data points. Here $\rho(h) = \text{Corr}[Y_t, Y_{t+h}]$ for $h = 1, 2$

a_1	−0.800	−0.500	−0.400	0.100	0.250	0.250
b_1	−0.430	−1.000	−0.350	0.200	0.550	0.730
$\rho(1)$	−0.984	−0.519	−0.188	0.145	0.630	0.979
$\rho(2)$	0.997	0.613	0.117	0.016	0.500	0.959

with

$$\zeta_t = \frac{Y_t - \lambda_t}{\lambda_t^{\delta}}, \tag{18}$$

where β_i, $i = 0, \ldots, p$ (with $\beta_i \neq 0$ for $i = 1, 2, \ldots, p$) are unknown regression parameters and $\delta \in (0, 1]$. If $\delta = 1/2$ then (17) is a moving average model of the so-called Pearson residuals; see definition (29). Under the above specification, we have the following results:

- The mean of the sequence $\{\zeta_t\}$ is zero:

$$\text{E}[\zeta_t] = 0.$$

- The variance of the sequence $\{\zeta_t\}$ is given by the following:

$$\text{Var}[\zeta_t] = \text{E}[\lambda_t^{1-2\delta}].$$

- The mean and ACF of the log-mean process $\{\nu_t\}$ are:

$$\text{E}[\nu_t] = \beta_0,$$

and

$$\text{Cov}[\nu_t, \nu_{t+h}] = \begin{cases} \displaystyle\sum_{i=1}^{p-h} \beta_i \beta_{i+h} \lambda_{t-i}^{1-2\delta}, & h \leq p, \\ 0, & \text{otherwise.} \end{cases}$$

We note that when $\delta = 1/2$, all the above expressions do not depend on t. In particular, the autocovariance function between ν_t and ν_{t+h}, for $h > 0$ reduces to the autocovariance functions of a standard moving average model of order p.

REMARK 1. As it was already mentioned, one of the advantages of model (14) is that time-dependent covariates can be easily introduced. To be more specific, suppose

that $\{X_t\}$ is some covariate time series. Then enlarging the σ-field to $\mathcal{F}_t^{Y,X,\lambda} = \sigma(Y_s, X_{s+1}, \lambda_0, s \le t)$ we obtain the model

$$Y_t \mid \mathcal{F}_{t-1}^{Y,X,\lambda} \sim \text{Poisson}(\lambda_t), \quad \nu_t = d + a_1\nu_{t-1} + b_1 \log(Y_{t-1} + 1) + cX_t, \quad t \ge 1, \tag{19}$$

where c is, in general, a real-valued parameter. Some remarks about the possible choices of the parameter c will be made in later sections. This remark, with obvious modifications, also applies to the case of model (17) as well. □

3.3. Nonlinear models for count time series

A large class of models for the analysis of count time series is given by the following specification

$$Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda}, \quad \lambda_t = f(\lambda_{t-1}, Y_{t-1}), \quad t \ge 1, \tag{20}$$

where $f(\cdot)$ is a known function up to an unknown finite dimensional parameter vector. Moreover, $f(\cdot)$ takes values on the positive real line, that is, $f : (0, \infty) \times \mathbb{N} \to (0, \infty)$ and the initial values Y_0 and λ_0 are assumed again to be random. An interesting example of a nonlinear regression model for count time series analysis is given by the following specification

$$f(\lambda, y) = d + (a_1 + c_1 \exp(-\gamma\lambda^2))\lambda + b_1 y, \tag{21}$$

where d, a_1, c_1, b_1, γ are positive parameters. The above model is rather similar to the traditional exponential autoregressive model, see Haggan and Ozaki (1981). In the study by Fokianos et al. (2009), model (21) was studied for the case $d = 0$. Note that the parameter γ introduces a perturbation of the linear model (10), in the sense that when γ tends either to 0 or infinity, then (21) approaches two distinct linear models. An obvious generalization of model (20) is given by the following specification of the mean process

$$\lambda_t = f(\lambda_{t-1}, \ldots, \lambda_{t-p}, Y_{t-1}, \ldots, Y_{t-q}), \tag{22}$$

where $f(.)$ is function such that $f : (0, \infty)^p \times \mathbb{N}^q \to (0, \infty)$. Such examples are provided by the class of smooth transition autoregressive models of which the exponential autoregressive model is a special case (cf. Teräsvirta et al., 2010). Further examples of nonlinear time series models can be found in the works of Tong (1990) and Fan and Yao (2003). These models have not been considered in the literature earlier in the context of generalized linear models for count time series, and they provide a flexible framework for studying dependent count data.

3.4. Inference

We illustrate conditional maximum likelihood inference for the linear model (10). The methodology is quite analogous for models (14) and (20), so it is omitted. However, for

models such as (21), the presence of nonlinear parameters requires larger sample sizes for accurate estimation. Recall now (10) and let $\boldsymbol{\theta}$ be the three dimensional vector of unknown parameters, that is, $\boldsymbol{\theta} = (d, a_1, b_1)^{'}$, and let the true value of the parameter be $\boldsymbol{\theta}_0 = (d_0, a_1;_0, b_{1,0})^{'}$. Then, the conditional likelihood function for $\boldsymbol{\theta}$ based on model (10) and given a starting value λ_0 is given by

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{n} \frac{\exp(-\lambda_t(\boldsymbol{\theta}))\lambda_t^{Y_t}(\boldsymbol{\theta})}{Y_t!}.$$

Here we use the Poisson assumption, $\lambda_t(\boldsymbol{\theta}) = d + a_1\lambda_{t-1}(\boldsymbol{\theta}) + b_1 Y_{t-1}$ by (10) and $\lambda_t = \lambda_t(\boldsymbol{\theta}_0)$. Hence, the log-likelihood function is given up to a constant, by

$$l(\boldsymbol{\theta}) = \sum_{t=1}^{n} l_t(\boldsymbol{\theta}) = \sum_{t=1}^{n} \left(Y_t \log \lambda_t(\boldsymbol{\theta}) - \lambda_t(\boldsymbol{\theta})\right), \tag{23}$$

and the score function is defined by

$$\boldsymbol{S}_n(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \left(\frac{Y_t}{\lambda_t(\boldsymbol{\theta})} - 1\right) \frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{24}$$

where $\partial\lambda_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is a three-dimensional vector with components given by

$$\begin{aligned} &\frac{\partial \lambda_t}{\partial d} = 1 + a_1 \frac{\partial \lambda_{t-1}}{\partial d}, \quad \frac{\partial \lambda_t}{\partial a_1} = \lambda_{t-1} + a_1 \frac{\partial \lambda_{t-1}}{\partial a_1}, \\ &\frac{\partial \lambda_t}{\partial b_1} = Y_{t-1} + a_1 \frac{\partial \lambda_{t-1}}{\partial b_1}. \end{aligned} \tag{25}$$

The solution of the equation $\boldsymbol{S}_n(\boldsymbol{\theta}) = 0$, if it exists, yields the conditional maximum likelihood estimator of $\boldsymbol{\theta}$, which is denoted by $\widehat{\boldsymbol{\theta}}$. Furthermore, the Hessian matrix for model (10) is obtained by further differentiation of the score equations (24),

$$\begin{aligned} \boldsymbol{H}_n(\boldsymbol{\theta}) &= -\sum_{t=1}^{n} \frac{\partial^2 l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{'}} \\ &= \sum_{t=1}^{n} \frac{Y_t}{\lambda_t^2(\boldsymbol{\theta})} \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{'} - \sum_{t=1}^{n} \left(\frac{Y_t}{\lambda_t(\boldsymbol{\theta})} - 1\right) \frac{\partial^2 \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{'}}. \end{aligned} \tag{26}$$

The conditional information matrix is defined by

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \mathrm{Var}\left[\frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mid \mathcal{F}_{t-1}^{Y,\lambda}\right] = \sum_{t=1}^{n} \frac{1}{\lambda_t(\boldsymbol{\theta})} \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{'}, \tag{27}$$

and plays a crucial role in the asymptotic distribution of the MLE $\widehat{\boldsymbol{\theta}}$. More specifically, under certain regularity conditions, it can be proved that $\widehat{\boldsymbol{\theta}}$ is consistent and asymptotically normal, i.e.,

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{D} \mathcal{N}(0, \boldsymbol{G}^{-1}),$$

with the matrix $\boldsymbol{G}$ defined by

$$\boldsymbol{G}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{1}{\lambda_t}\left(\frac{\partial \lambda_t}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \lambda_t}{\partial \boldsymbol{\theta}}\right)'\right],$$

where $\mathrm{E}[\cdot]$ is taken with respect to the stationary distribution. All the above quantities can be computed, and they are employed for constructing predictions, confidence intervals, and so on. Although, the above formulaes are given for the linear model (10), they can be modified suitably for the log-linear model (14) and the nonlinear model (20).

3.5. *On the asymptotic distribution of the MLE*

It can be proved for regression models of the form (10), (14), and more generally for models like (20), that the MLE $\widehat{\boldsymbol{\theta}}$ is asymptotically normally distributed, as it was mentioned before. This is an important fact, since inference is based on this approximation. However, to study the asymptotic theory there is need to develop a central limit theory for the bivariate process $\{(Y_t, \lambda_t)\}$. We mention the approach taken by Neumann (2011), who studies model (20) and shows that the bivariate process $\{(Y_t, \lambda_t)\}$ has a unique stationary distribution and the response process is absolutely regular. In addition, Franke (2010) considers (22) and shows that the response process is weakly dependent with finite first moment; see Doukhan and Louhichi (1999) and the recent monograph by Dedecker et al. (2007) for definition of weak dependence and further examples. Using the general model (22), the essential condition assumed by both the above references, is that the function $f(\cdot)$ is a contraction, that is, for any $(\lambda_1, \ldots, \lambda_p, y_1, \ldots, y_q)$ and $(\lambda'_1, \ldots, \lambda'_p, y'_1, \ldots, y'_q)$

$$\begin{aligned} &|f(\lambda_1, \ldots, \lambda_p, y_1, \ldots, y_q) - f((\lambda'_1, \ldots, \lambda'_p, y'_1, \ldots, y'_q)| \\ &\quad \leq \sum_{i=1}^{p} \alpha_i |\lambda_i - \lambda'_i| + \sum_{j=1}^{q} \gamma_j |y_j - y'_j|, \end{aligned} \tag{28}$$

where $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \gamma_j < 1$. This is the same condition assumed by Fokianos et al. (2009) and Fokianos and Tjøstheim (2012), whose approach is based on Markov chains theory.

Turning now to the questions regarding ergodicity and inference, we note that these problems have been examined in detail by Fokianos et al. (2009), Fokianos and Tjøstheim (2011, 2012) (see also Woodard et al. (2011)), who also use a perturbation argument to prove geometric ergodicity of $\{(Y_t, \lambda_t)\}$. This means that instead of proving geometric ergodicity of $\{(Y_t, \lambda_t)\}$, the authors are considering a perturbed $\{(Y_t^m, \lambda_t^m, U_t)\}$, where $\{U_t\}$ is a sequence of i.i.d uniform random variables. The strategy to study the properties of the bivariate process $\{(Y_t, \lambda_t)\}$ is to prove geometric ergodicity of $\{(Y_t^m, \lambda_t^m, U_t)\}$ and then to use this fact to obtain asymptotic normality for the likelihood estimators. Asymptotic normality of the likelihood estimates of the nonperturbed model is proved by employing an approximation lemma, which gives conditions for the proximity of the perturbed version to nonperturbed version. Detailed exposition

of the perturbation argument can be found in the references mentioned above. Here, we list the following up-to-date known facts for models (10) and (14):

1. For the linear model (10):
   (a) Consider the perturbed linear model and suppose that $0<a_1+b_1<1$. Then, the process $\{(Y_t^m, \lambda_t^m, U_t), t \geq 0\}$ is a $V_{(Y,U,\lambda)}$-geometrically ergodic Markov chain with $V_{Y,U,\lambda}(Y,U,\lambda) = 1 + Y^k + \lambda^k + U^k$.
   (b) If $0 < a_1 + b_1 < 1$, then the perturbed model can be made arbitrarily close to the unperturbed model.
   (c) If $0 < a_1 + b_1 < 1$, then the conditional maximum likelihood estimators of (d, a_1, b_1) are consistent and asymptotically normally distributed.
2. For the log-linear model (14), define $\{(Y_t^m, \nu_t^m, U_t), t \geq 0\}$ as its perturbed version.
   (a) Suppose that $\mid a_1 \mid < 1$. In addition, assume that when $b_1 > 0$, then $|a_1 + b_1| < 1$, and when $b_1 < 0$, then $|a_1||a_1 + b_1| < 1$. Then, the process $\{(Y_t^m, U_t, \nu_t^m), t \geq 0\}$ is a $V_{(Y,U,\nu)}$-geometrically ergodic Markov chain with $V_{Y,U,\lambda}(Y,U,\nu) = 1 + \log^{2k}(1+Y) + \nu^{2k} + U^{2k}$, k being a positive integer.
   (b) If $|a_1 + b_1| < 1$, whenever a_1 and b_1 have the same sign, and $a_1^2 + b_1^2 < 1$ whenever a_1 and b_1 have different signs, then the perturbed log-linear model can be made arbitrarily close to the unperturbed log-linear model.
   (c) If $|a_1 + b_1| < 1$, whenever a_1 and b_1 have the same sign, and $a_1^2 + b_1^2 < 1$, whenever a_1 and b_1 have different signs, then the conditional maximum likelihood estimators of (d, a_1, b_1) are consistent and asymptotically normally distributed.

We clarify the above results by considering the first statement about the linear model (10). Result 1(a) implies that when $0 < a_1 + b_1 < 1$, then the perturbed model possesses moments of any order and any average of functions of $\{(Y_t^m, \lambda_t^m, U_t), t \geq 0\}$ will converge weakly to its expected value. This fact has important consequences, because it allows the study of the maximum likelihood estimators derived by (23) given that the unperturbed model is close to the perturbed model under the same condition. For the log-linear model (14), the conditions for proving that the perturbed model approaches the unperturbed model are quite restrictive when compared to the conditions for geometric ergodicity. The same phenomenon occurs for model (17) for the case $p = 1$; see Davis et al. (2005), who prove asymptotic normality of the maximum likelihood estimators when $\delta = 1$ and $\beta_1 > 0$, such that $\beta_1(1 + \exp(\beta_1 - \beta_0))^{1/2} < 1$. However, it was shown by Davis et al. (2003) that if $1/2 \leq \delta \leq 1$, then the chain $\{\nu_t\}$ has a stationary distribution. In particular, when $\delta = 1$, then $\{\nu_t\}$ is uniformly ergodic and has a unique stationary distribution.

To complement the presentation, ergodicity of model (20) has been proved by employing the contraction assumption (28) for $p = q = 1$ on $(\cdot)$, Fokianos and Tjøstheim (2012) for its perturbed version, and Neumann (2011) and Franke (2010) for the response process. Under such assumption Fokianos and Tjøstheim (2012) show the asymptotic normality of the MLE for mode (20) showing that the perturbed version approximates the nonperturbed version. We close this part by the following important remarks.

Remark 2. For a log-linear model which includes covariates, such as (19), the estimation problem is attacked along the lines described in Section 3.4. To study ergodicity and asymptotic normality of the MLE in this case, suppose that $\{X_t\}$ a real-valued

Markov chain which possess a density. Then, we can construct a two-dimensional Markov chain $\{\nu_t, X_{t+1}\}$ and a corresponding three-dimensional chain with $\{Y_t\}$ included. If the transition mechanism of $\{X_t\}$ does not depend on $\{\nu_t, Y_t\}$, it is simple to find conditions for geometric ergodicity; see Fokianos and Tjøstheim (2011), for more details. □

Remark 3. We note, however, that the asymptotic theory concerning the maximum likelihood estimators for the regression parameters has been developed under the assumption of the Poisson distribution. Such an approach poses several robustness issues related to model misspecification. A possible venue to overcome this problem is the quasi-likelihood estimation method, see Heyde (1997) and Kedem and Fokianos (2002, Section 1.7), for instance. In this case, the score is determined by a mean regression equation and a working variance function. Such methods have been explored, for example, in the GARCH framework by Berkes et al. (2003) and it is worth studying their performance in the context of count time series regression models. □

3.6. *Data examples*

The above theory is applied to the real data examples discussed in the Introduction; recall Fig. 1. For both time series, the mean is always less than their variance. In other words, the data exhibits overdispersion – a fact that holds for all the Poisson-distributed models that were discussed so far. For the analysis of those time series, we fit both the linear model (10) and the log-linear model (14). To model these data, set $\lambda_0 = 0$ and $\partial\lambda_0/\partial\boldsymbol{\theta} = 0$ for initialization of the recursions in the case of the linear model; see Eqs (25). For the log-linear model, the corresponding initializations are set to $\nu_0 = 1$ and $\partial\nu_0/\partial\boldsymbol{\theta} = \mathbf{0}$. Table 2 lists the results of the analysis. The numbers in parentheses, next to the estimators, correspond to the standard errors of the estimates. These are computed by using the so-called robust sandwich matrix $H_n(\hat{\theta})G_n^{-1}(\hat{\theta})H_n(\hat{\theta})$, where $G_n(\hat{\theta})$ has been defined by (27) and $H_n(\theta)$ is given by (26). To examine the adequacy of the fit, consider the so-called Pearson residuals (recall (18) with $\delta = 1/2$)

$$e_t = \frac{Y_t - \lambda_t}{\sqrt{\lambda_t}}, \quad t \geq 1. \tag{29}$$

Table 2
Data analysis results

Linear Model				Log-linear Model Fit			
Series C3							
$\hat{d}$	$\hat{a}_1$	$\hat{b}_1$	MSE	$\hat{d}$	$\hat{a}_1$	$\hat{b}_1$	MSE
2.385 (0.533)	0.050 (0.088)	0.5603 (0.073)	1.285	0.476 (0.183)	0.080 (0.097)	0.619 (0.084)	1.296
Transactions Data							
$\hat{d}$	$\hat{a}_1$	$\hat{b}_1$	MSE	$\hat{d}$	$\hat{a}_1$	$\hat{b}_1$	MSE
0.581 (0.162)	0.744 (0.026)	0.198 (0.016)	2.367	0.105 (0.034)	0.746 (0.026)	0.207 (0.019)	2.391

Under the true model, the process $\{e_t\}$ is a white noise sequence with constant variance; see Kedem and Fokianos (2002, Section 1.6.3). To estimate the Pearson residuals, substitute λ_t by $\hat{\lambda}_t \equiv \lambda_t(\hat{\boldsymbol{\theta}})$. Comparison among the models is implemented by calculating the mean square error (MSE) of the Pearson residuals, which is given by $\sum_{t=1}^{N} \hat{e}_t^2/(N-p)$, where p denotes the number of estimated parameters; see Kedem and Fokianos (2002, Section 1.8) for more details on diagnostics (see also Zhu and Wang (2010), for a recent contribution directly related to models of the form (6)).

Table 2 summarizes the findings of the data analysis. Consider first the C3 series. We note that both linear and log-linear models yield almost the same MSE and the estimators obtained for a_1 and b_1 are similar from both models. In fact, the feedback mechanism does not provide any improvement for the fit, because the estimator of a_1 is large when compared to its standard error, for both models. Figure 4 shows the results of the data analysis that point to the adequacy of the fit. Note that the bottom plot shows the cumulative periodogram plot of the Pearson residuals, which confirms that the sequence (29) is white noise. One practical aspect that arises in applications of model (14) is the choice of $\log(Y_{t-1}+1)$ in the regression equation. Here, we mention that in order to examine the sensitivity of the results, as a function of the log term in model (14), we can work as follows. Fit the following series of models to the log-mean

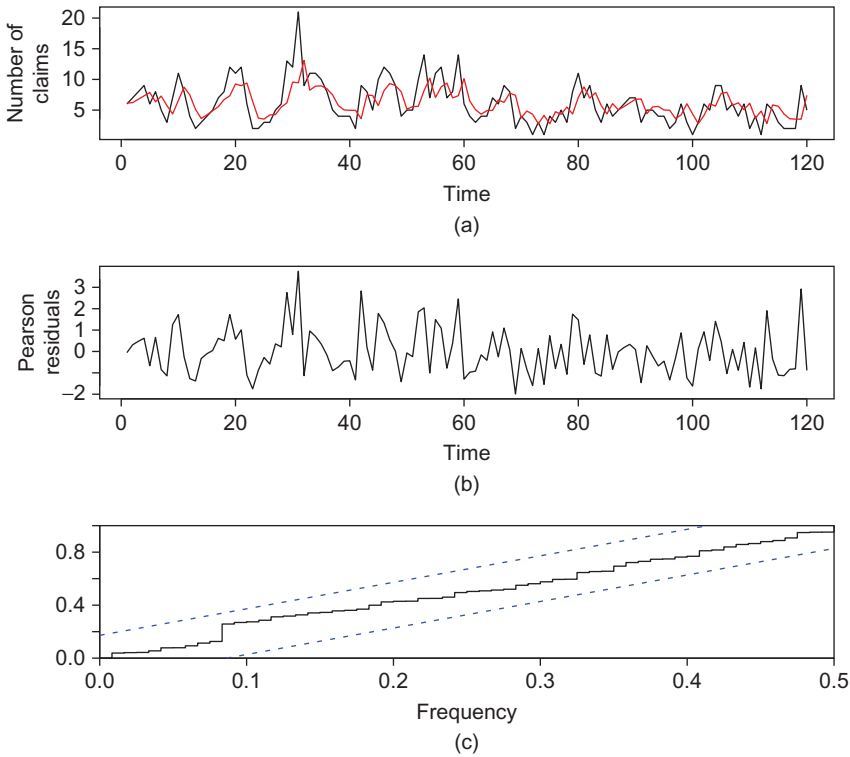

Fig. 4. (a) Series C3. The red line corresponds to the prediction $\lambda_t(\hat{\boldsymbol{\theta}})$ obtained by fitting model (14). (b) Pearson Residuals obtained by fitting model (14). (c) Cumulative periodogram plot of the Pearson residuals.

processes

$$v_t = d + a_1 v_{t-1} + b_1 \log(Y_{t-1} + v),$$

for both time series, where v is a constant that takes values from 1 to 10 (or some other bound) with step equal to 0.5. Then, calculate the MSE of the Pearson residuals for all different model specifications obtained by varying the constant v and compare them. For the C3 series, the sample variance of obtained MSE values is almost zero. In conclusion, we see that the choice of $\log(Y_{t-1} + 1)$ does not affect the results of the analysis greatly, at least for the C3 series.

Turning now to the transactions data, we see again that both the models (10) and (14) yield similar MSE values. Note that the sum of estimated coefficients is close to one for both linear and log-linear model. This corresponds to a frequently observed phenomenon for GARCH(1,1) models. Figure 5 demonstrates again the adequacy of the fit for the log-linear model. To examine the sensitivity of the results as a function of the log term in model (14), we repeat the previous exercise with a constant v that takes values from 1 to 10 with step equal to 0.5. The MSE values have a range between 2.389 and 2.391. Therefore, we observe again that choice of $\log(Y_{t-1} + 1)$ does not affect the results of the analysis greatly.

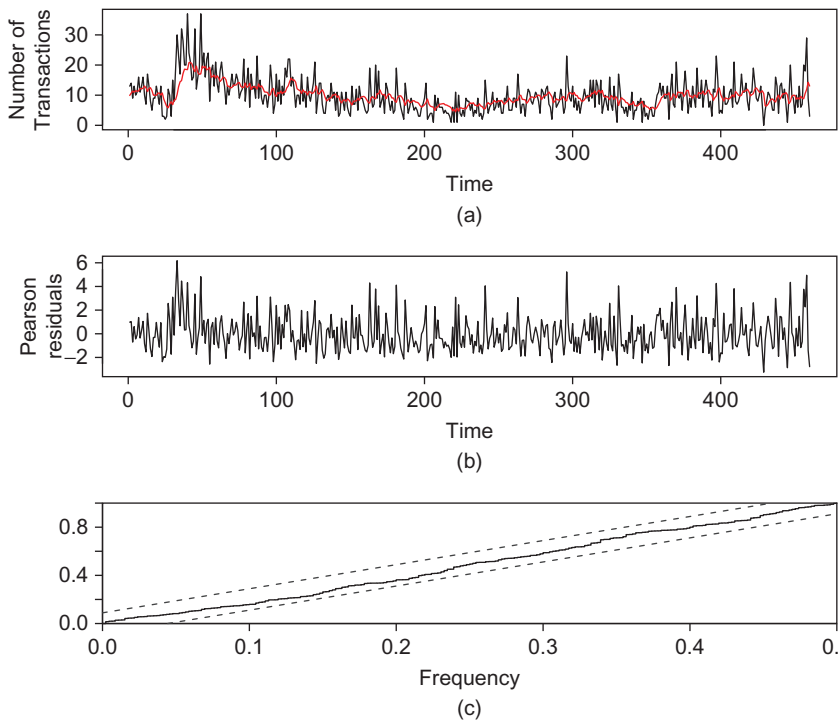

Fig. 5. (a) Transactions data. The red line corresponds to the prediction $\lambda_t(\hat{\boldsymbol{\theta}})$ obtained by fitting model (14). (b) Pearson Residuals obtained by fitting model (14). (c) Cumulative periodogram plot of the Pearson residuals.

As a general remark, models like (10) and (14) will be more useful when applied to count time series data with strong correlation. The feedback mechanism allows for more parsimonious modeling. This is in accordance with the GARCH methodology, whereby lagged values of volatility allow for parsimony. Furthermore, when the data are positively correlated, then models (10) and (14) will yield similar conclusions. It is anticipated that the log-linear model (14) provides a better fit when either there exists negative correlation among the data or when covariates need to be taken into account for the data analysis.

4. Other regression models for count time series

The Poisson distribution is the most natural candidate among discrete distributions to model count data. However, the literature offers several alternatives to Poisson. In this section, we discuss the case of negative binomial distribution and the double Poisson distribution as alternative models for the analysis of count time series. We also survey other alternative regression-based methods for count time series analysis.

4.1. Other distributional assumptions

Recall that if Y is random variable that follows the negative binomial distribution with parameters (r, θ), where $\theta \in (0, 1)$ and r an integer, then its probability mass function is given by

$$\mathrm{P}[Y = y] = \binom{y + r - 1}{y} \theta^y (1 - \theta)^r, \quad y = 0, 1, 2, \ldots. \tag{30}$$

Accordingly, we denote $Y \sim \mathrm{NegBin}(r, \theta)$. With this notation, it is well known that $\mathrm{E}[Y] = r\theta/(1 - \theta)$ and $\mathrm{Var}[Y] = r\theta/(1 - \theta)^2$.

Consider again $\{Y_t\}$ to be the response and assume the following model, see Zhu (2011):

$$Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda} \sim \mathrm{NegBin}(r, \theta_t), \quad \lambda_t \equiv \frac{\theta_t}{1 - \theta_t} = d + a_1 \lambda_{t-1} + b_1 Y_{t-1}, \quad t \geq 1, \tag{31}$$

where the parameters d, a_1, b_1 are all non-negative and λ_0, Y_0 are some random starting values. The above model regresses the log-odds of θ_t to its past values and past values of the responses. More generally, we can study models of the form

$$\lambda_t = d + \sum_{i=1}^{p} a_i \lambda_{t-i} + \sum_{j=1}^{q} b_j Y_{t-j}, \quad t \geq \max(p, q),$$

but we will insist on the simpler model (31) for ease of presentation (see Zhu (2011) for more details). With the same notation as before, it is easily seen that this particular specification yields again

$$\mathrm{E}[Y_t] = \mathrm{E}\Big(\mathrm{E}\Big[Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda}\Big]\Big) = r\mathrm{E}[\lambda_t].$$

Therefore, assuming stationarity, we obtain, from (31), that

$$\mathrm{E}[Y_t] = r\frac{d}{1-a_1-rb_1},$$

provided that $a_1 + rb_1 < 1$. We will be working again as in the case of model (10) in order to understand the dynamics of model (31). Toward this goal, consider again the following representation

$$Y_t = r\lambda_t + (Y_t - r\lambda_t) = rd + ra_1\lambda_{t-1} + rb_1 Y_{t-1} + \epsilon_t, \tag{32}$$

where the error term $\{\epsilon_t\}$ is again a white noise sequence. This fact is proved next, by assuming the condition $a_1 + rb_1 < 1$. Furthermore, Eq. (32) implies that the observed process depends on its past values and on its past odds of the sequence of probabilities $\{\theta_t\}$. The details for proving that the sequence $\{\epsilon_t\}$ is white noise are as follows:

- Constant mean:

$$\mathrm{E}[\epsilon_t] = \mathrm{E}\Big[(Y_t - r\lambda_t)\Big] = \mathrm{E}\Big[\mathrm{E}\Big(Y_t - r\lambda_t \mid \mathcal{F}_{t-1}^{Y,\lambda}\Big)\Big] = 0,$$

- Constant variance:

$$\begin{aligned}\mathrm{Var}[\epsilon_t] &= \mathrm{Var}\Big[\mathrm{E}\Big(\epsilon_t \mid \mathcal{F}_{t-1}^{Y,\lambda}\Big)\Big] + \mathrm{E}\Big[\mathrm{Var}\Big(\epsilon_t \mid \mathcal{F}_{t-1}^{Y,\lambda}\Big)\Big] = r\mathrm{E}[\lambda_t(1+\lambda_t)] \\ &= \frac{1-(a_1+rb_1)^2}{1-(a_1+rb_1)^2-rb_1^2}\Big(\mathrm{E}[Y_t] + \frac{\mathrm{E}^2[Y_t]}{r}\Big),\end{aligned} \tag{33}$$

  which is independent of t, since $\{Y_t\}$ is stationary. Equation (33) is proved in the Appendix.
- Uncorrelated sequence: For $k > 0$,

$$\mathrm{Cov}(\epsilon_t, \epsilon_{t+k}) = \mathrm{E}[\epsilon_t\epsilon_{t+k}] = \mathrm{E}\Big[\epsilon_t\mathrm{E}\Big(\epsilon_{t+k} \mid \mathcal{F}_{t+k-1}^{Y,\lambda}\Big)\Big] = 0$$

To study the second-order properties of (31), we employ representation (32) using the same technique as that which was employed for deriving (12). More specifically, Eq. (32) shows that the $\{Y_t\}$ process can be expressed as

$$\Big(Y_t - \frac{rd}{1-a_1-rb_1}\Big) = (a_1+rb_1)\Big(Y_{t-1} - \frac{rd}{1-a_1-rb_1}\Big) + \epsilon_t - a_1\epsilon_{t-1}. \tag{34}$$

This is an ARMA(1,1) process and therefore when $0 < (a_1+rb_1)^2 + rb_1^2 < 1$ we have that $\{Y_t\}$ is second order stationary with autocovariance function

$$\mathrm{Cov}\,[Y_t, Y_{t+h}] = \begin{cases} \dfrac{(1-(a_1+rb_1)^2+r^2b_1^2)}{1-(a_1+rb_1)^2-rb_1^2}\Big(\mathrm{E}[Y_t] + \dfrac{\mathrm{E}^2[Y_t]}{r}\Big), & h = 0, \\ \dfrac{rb_1(1-a_1(a_1+rb_1))}{1-(a_1+rb_1)^2-rb_1^2}\Big(\mathrm{E}[Y_t] + \dfrac{\mathrm{E}^2[Y_t]}{r}\Big)(a_1+rb_1)^{h-1}, & h \geq 1. \end{cases}$$

It is clear that the ACF of model (31) is equal to (see Appendix)

$$\text{Corr}\,[Y_t, Y_{t+h}] = \frac{rb_1(1-a_1(a_1+rb_1))}{1-(a_1+rb_1)^2+r^2b_1^2}(a_1+rb_1)^{h-1}, \quad h \geq 1. \tag{35}$$

Estimation for model (31) is based on the maximum likelihood method, where a profiling procedure is employed. For a grid of values of r (recall that r is positive integer) the negative binomial log-likelihood function is maximized with respect to (d, a_1, b_1). Then, we estimate r by the value that maximizes all log-likelihood functions. For this choice of r, the regression parameters are estimated. This method implies that the standard errors of the parameter estimators need to be calculated by resampling methods because the methodology corresponds to a two-stage procedure. The problem of estimation of the parameter r is challenging and the interpretation of its value is unclear. A better way to deal with this issue is to define the negative binomial probability mass function (30) by

$$\text{P}[Y=y] = \frac{\Gamma(y+k)}{y!\,\Gamma(k)}\left(\frac{k}{\lambda+k}\right)^k\left(\frac{\lambda}{\lambda+k}\right)^y, \quad y=0,1,2,\ldots,$$

where $k > 0$. This is a plain consequence of the fact that the negative binomial distribution is a mixture of Poisson random variables. Then we employ model (31) – or its generalization – with obvious modifications.

An alternative way to relax the Poisson distributional assumption is given by the double Poisson distribution (Efron, 1986). The double Poisson distribution is an exponential combination of two Poisson densities, that is

$$f(y\,;\lambda,\theta) = C(\lambda,\theta)\,[\text{Poisson}(\lambda)]^{\theta}\,[\text{Poisson}(y)]^{1-\theta},$$

where θ is a dispersion parameter and $C(\lambda,\theta)$ is the normalizing constant. It can be shown that

$$\frac{1}{C(\lambda,\theta)} \approx 1+\frac{1-\theta}{12\theta\lambda}\left(1+\frac{1}{\theta\lambda}\right),$$

and that the mean and variance of the double Poisson distribution are approximately equal to λ and λ/θ, respectively. For the Double Poisson model, we can use models such as (10) to model the mean process, see Kedem and Fokianos (2002, Section 4.6, Problem 4) and Heinen (2003). Properties of maximum likelihood estimators derived by imposing either the negative binomial distribution or the double Poisson distribution is a research topic that has not been addressed suitably in the literature. As a final remark, we note that several other alternative distributional assumptions can be adopted along the previous lines; for example, data can be modeled by means of the zero-inflated Poisson model (Lambert, 1992), or the truncated Poisson model (Fokianos, 2001), and so on. However, the likely gains of such approaches will depend, in general, upon the context of their application.

4.2. Parameter driven models

So far we have discussed models that fall under the framework of observation driven model. This implies that even though the mean process $\{\lambda_t\}$ is not observed directly, it can still be recovered explicitly as function of the past responses, see Eq. (13) for example. However, a different point of view has been taken by Zeger (1988), who introduced a regression models for time series of counts by assuming that the observed process is driven by a latent (unobserved) process. To be more specific suppose that, conditional on an unobserved process $\{\xi_t, t \geq 1\}$, $\{Y_t, t \geq 1\}$, is a sequence of independent counts such that

$$\mathrm{E}[Y_t \mid \xi_t] = \mathrm{Var}[Y_t \mid \xi_t] = \xi_t \exp(d + a_1 y_{t-1}). \tag{36}$$

In the above we consider a simple model for illustration, but more complex models that include higher order lagged values of the response and any covariates can be included in (36). Assume that $\{\xi_t\}$ is a stationary process with $\mathrm{E}[\xi_t] = 1$ and $\mathrm{Cov}[\xi_t, \xi_{t+h}] = \sigma^2 \rho_\xi(h)$, for $h \geq 0$. Then, it can be proved that

$$\mathrm{E}[Y_t] = \mathrm{E}[\exp(d + a_1 Y_{t-1})], \quad \mathrm{Cov}[Y_t, Y_{t+h}] = \sigma^2 \mathrm{E}[Y_t]\mathrm{E}[Y_{t+h}]\rho_\xi(h).$$

It is clear that the above formulation, although similar to a Poisson–loglinear model, reveals that the observed data are overdispersed. Estimation of all unknown parameters is discussed by Zeger (1988). Further detailed study of model (36) can be found in the study by Davis et al. (2000), where the authors address the problem of existence of the latent stochastic process $\{\xi_t\}$ and derive the asymptotic distribution of the regression coefficients when the latter exist. They also suggest adjustments for the estimators of σ^2 and of the autocovariance. In the context of negative binomial regression, the latent process model (36) has been extended by Davis and Wu (2009). See also Harvey and Fernandes (1989) for a state-space approach with conjugate priors for the analysis of count time series and Jørgensen et al. (1999) for multivariate count longitudinal data. More generally, state-space models for count time series are discussed by West and Harrison (1997), Durbin and Koopman (2001), Cappé et al. (2005), and among others.

5. Integer autoregressive models

We discuss another class of models for integer-valued time series. This class consists of the so-called integer autoregressive models that are constructed by means of the thinning operator. As we shall see, these models can be viewed as a special case of a branching process with immigration; see Kedem and Fokianos (2002, Chapter 5) for a detailed account of integer AR and MA processes. The notation $\{Y_t\}$ still refers to the response process.

5.1. Branching processes

An important model for integer-valued time series is the branching process with immigration, also known as the Galton–Watson process with immigration. It is

defined by

$$Y_t = \sum_{i=1}^{Y_{t-1}} X_{t,i} + I_t, \quad t = 1, \ldots, \tag{37}$$

where the initial value Y_0 is a non-negative integer-valued random variable, and $\sum_1^0 \equiv 0$. The processes $\{X_{t,i}\}$ and $\{I_t\}$ drive the dynamics of the system and they are mutually independent, independent of Y_0, and each consisting of i.i.d. random variables. This defines a Markov chain $\{Y_t\}$ with non-negative integer states. Model (37) was originally introduced and applied by Smoluchowski in 1916 for studying the fluctuations in the number of particles contained in a small volume in connection with the second law of thermodynamics; see Chandrasekhar (1943). Since then, the process has been applied extensively in biological, sociological, and physical branching phenomena, see for instance Kedem and Chiu (1987), Franke and Seligmann (1993), Berglund and Brännäs (2001), Böckenholt (1999), and the review by McKenzie (2003), Weiß (2008), and Jung and Tremayne (2011).

Note that Y_t is the size of the t'th generation of a population, $X_{t,1}, \ldots, Y_{t,Y_{t-1}}$ are the offspring of the $(t-1)$st generation, and I_t is the contribution of immigration to the t'th generation. An important role in the behavior of $\{Y_t\}$ is played by the mean $m = \mathrm{E}[X_{t,i}]$ of the offspring distribution, where the cases $m < 1, m = 1, m > 1$, are referred to as subcritical, critical, and supercritical, respectively. In the subcritical case $\{Y_t\}$ has a limiting stationary distribution, whereas in the supercritical case $\{Y_t\}$ explodes at an exponential rate. In the critical case, the process is either null recurrent or transient.

The process (37) admits a useful autoregressive representation, similar to the model representation (32), for instance. Let $\lambda = \mathrm{E}[I_t]$, and let $\mathcal{F}_t$ be generated by the past information $Y_0, Y_1, Y_2, \ldots, Y_t$. Then $\mathrm{E}[Y_t \mid \mathcal{F}_{t-1}] = mY_{t-1} + \lambda$. Therefore, with $\epsilon_t \equiv Y_t - \mathrm{E}[Y_t|\mathcal{F}_{t-1}]$, the stochastic Eq. (37) is transformed into a stochastic regression model,

$$Y_t = mY_{t-1} + \lambda + \epsilon_t, \quad t = 1, \ldots, \tag{38}$$

as before. The noise process $\{\epsilon_t\}$ consists of uncorrelated random variables such that $\mathrm{E}[\epsilon_t] = 0$. However, $\mathrm{E}[\epsilon_t^2 \mid \mathcal{F}_{t-1}] = \mathrm{Var}[X_{t,i}]Y_{t-1} + \mathrm{Var}[I_t]$ is unbounded as Y_{t-1} increases.

As suggested by (38), the least squares estimators for m, λ are obtained by minimizing,

$$\sum_{t=1}^{n} \epsilon_t^2 = \sum_{t=1}^{n} (Y_t - mY_{t-1} - \lambda)^2,$$

and are given by

$$\tilde{m} = \frac{\sum Y_t \sum Y_{t-1} - n \sum Y_t Y_{t-1}}{(\sum Y_{t-1})^2 - n \sum Y_{t-1}^2}$$

$$\tilde{\lambda} = \frac{\sum Y_{t-1} Y_t \sum Y_{t-1} - \sum Y_{t-1}^2 \sum Y_t}{(\sum Y_{t-1})^2 - n \sum Y_{t-1}^2},$$

where the summation limits are from $t = 1$ to $t = n$. It turns out that $\tilde{m}$ is consistent in all the three cases, whereas $\tilde{\lambda}$ is not consistent in the critical and supercritical cases. Improved estimators are obtained by weighted least squares. We write (38) as

$$\frac{Y_t}{\sqrt{Y_{t-1}+1}} = m\sqrt{Y_{t-1}+1} + \frac{(\lambda - m)}{\sqrt{Y_{t-1}+1}} + \frac{\epsilon_t}{\sqrt{Y_{t-1}+1}}, \tag{39}$$

and estimate m and $\lambda - m$ by minimizing $\sum \delta_t^2$ where $\delta_t = \epsilon_t/\sqrt{Y_{t-1}+1}$ to obtain (see Winnicki (1986)),

$$\hat{m} = \frac{\sum Y_t \sum \frac{1}{Y_{t-1}+1} - n \sum \frac{Y_t}{Y_{t-1}+1}}{\sum (Y_{t-1}+1) \sum \frac{1}{Y_{t-1}+1} - n^2}, \tag{40}$$

$$\hat{\lambda} = \frac{\sum Y_{t-1} \sum \frac{Y_t}{Y_{t-1}+1} - \sum Y_t \sum \frac{Y_{t-1}}{Y_{t-1}+1}}{\sum (Y_{t-1}+1) \sum \frac{1}{Y_{t-1}+1} - n^2}, \tag{41}$$

where again the summation limits are from 1 to n. Then for $0 < m < \infty$, $\hat{m} \to m$ in probability. That is $\hat{m}$ is consistent in all cases, provided that $m > 0$. Furthermore, the limiting distribution of $\hat{m}$ is normal in noncritical cases and non-normal in the critical case. On the other hand, $\hat{\lambda}$ is consistent for $m \leq 1$, but not for $m > 1$, and is asymptotically normal when $m < 1$ or $m = 1$ and $2\lambda > \mathrm{Var}[Y_{n,i}]$ (Wei and Winnicki, 1990; Winnicki, 1986).

5.2. *Thinning operator-based models*

In this section, we review models that are based on the thinning operator. The thinning operator is defined as follows (see Steutel and van Harn (1979)). Suppose that Y is a non-negative integer random variable and let $\alpha \in [0, 1]$. Then, the thinning operator, denoted by $\circ$, is defined as

$$\alpha \circ Y = \sum_{i=1}^{Y} X_i,$$

where $\{X_i\}$ is a sequence of i.i.d. Bernoulli random variables – independent of Y – with success probability α. The sequence $\{X_i\}$ is termed as counting series. The random variable $\alpha \circ Y$ counts the number of successes in a random number of Bernoulli trials, where the probability of success α remains constant throughout the experiment. Therefore, given $Y = y$, the random variable $\alpha \circ Y$ follows the binomial distribution with parameters y and α.

It turns out that the thinning operator is quite useful for modeling count time series. Building a model for count time series is based on a typical autoregressive model where scalar multiplication is replaced by thinning operators, see McKenzie (1985, 1986, 1988), Al-Osh and Alzaid (1987), Alzaid and Al-Osh (1990), and Du and Li (1991). Let us consider the simple integer autoregressive model of order 1, which is abbreviated by INAR(1). The INAR(1) model is a special case of the branching process

with immigration (37). However, it deserves special consideration due to the thinning operation or calculus. Suppose that $a_1 \in (0,1)$ and let $\{\epsilon_t\}$ be a sequence of i.i.d non-negative integer-valued random variables with $\mathrm{E}[\epsilon_t] = \mu$ and $\mathrm{Var}[\epsilon_t] = \sigma^2$. The integer autoregressive process of order 1, $\{Y_t, t \geq 1\}$, is defined as

$$Y_t = a_1 \circ Y_{t-1} + \epsilon_t, \quad t \geq 1, \tag{42}$$

where $a_1 \circ Y_{t-1}$ is the sum of Y_{t-1} Bernoulli random variables all of which are independent of Y_{t-1}. It should be noted that the Bernoulli variables used in $a_1 \circ Y_{t-1}$ are independent of those used in $a_1 \circ Y_{t-2}$, and so on. This is the assumption imposed by Du and Li (1991) and used subsequently in the majority of all published work related to integer autoregressive processes. Clearly, (42) is a special case of (37). Employing the same techniques that we used before (or by repeated substitution into (42) and use of the properties of the thinning operator), we obtain that the mean, variance, and ACF of the INAR(1) are given by

$$\mathrm{E}[Y_t] = \frac{\mu}{1-a_1}, \quad \mathrm{Var}[Y_t] = \frac{a_1\mu + \sigma^2}{1-a_1^2}, \quad \mathrm{Cov}[Y_t, Y_{t+h}] = a_1^h, \quad h \geq 1. \tag{43}$$

Note that the ACF decays exponentially with the lag h as in AR(1) models, but unlike the autocorrelation of a stationary AR(1) process, it is always positive for $a_1 \in (0,1)$. Furthermore, under suitable conditions, it can be shown that Y_t has a discrete self-decomposable distribution. This, in turn, implies unimodality properties and characterization of the distribution of Y_t through the sequence $\{\epsilon_t\}$. For instance, we can obtain the result that Y_t follows the Poisson distribution if and only if ϵ_t follows the Poisson distribution; see Al-Osh and Alzaid (1987).

Estimation in INAR(1) means estimation in the branching process with immigration in the subcritical case, and this has already been discussed earlier. Still it is interesting to note a few facts regarding estimation in the Poisson INAR(1). Estimation procedures for the parameters a_1 and μ of the INAR(1) model (42) assuming that the sequence $\{\epsilon_t\}$ follows the Poisson distribution, has been discussed by Al-Osh and Alzaid (1987). Imposing the Poisson assumption on the distribution of the error sequence $\{\epsilon_t\}$, and employing Eq. (43) yields a method of moments estimators for a_1 and μ (in this case $\mu = \sigma^2$), given by

$$\hat{a}_1 = \frac{\sum_{t=0}^{n-1}(Y_t - \bar{Y})(Y_{t+1} - \bar{Y})}{\sum_{t=0}^{N}(Y_t - \bar{Y})^2}, \quad \hat{\mu} = \frac{1}{n}\sum_{t=1}^{n}\hat{\epsilon}_t,$$

where $\hat{\epsilon}_t = Y_t - \hat{a}_1 Y_{t-1}$, for $t = 1, \ldots, n$. Alternatively, we can consider the conditional least squares of the parameters a_1 and μ, i.e., the values that minimize the residual sum of squares

$$\sum_{t=1}^{N}(X_t - a_1 X_{t-1} - \mu)^2 .$$

Asymptotic properties of the resulting estimators are deduced by using classical results from Klimko and Nelson (1978). In the works of Ispány et al. (2003), the authors study the INAR(1) model where the autoregressive coefficient $a_1 = a_{1n}$ satisfies $a_{1n} = 1 - \gamma_n/n$ with $\gamma_n \to \gamma > 0$. Such a sequence is called nearly unstable. The authors show that it can be approximated, in an appropriate sense, by a Gaussian martingale and use this result to show that the conditional least squares estimator of a_1 is asymptotically normal with the rate of convergence $n^{3/2}$. As a final remark, we note that application of maximum likelihood estimation requires a full distributional assumption about the innovations. With the Poisson assumption, the likelihood function of a time series $Y_0, Y_1, \ldots, Y_n$ from model (42) is

$$\left(\prod_{t=1}^{N} P_t(Y_t)\right) \frac{(\mu/(1-a_1))^{Y_0}}{Y_0!} \exp\left(-\mu/(1-a_1)\right),$$

$$P_t(y) = \exp(-\mu) \sum_{i=0}^{\min(Y_t, Y_{t-1})} \frac{\mu^{y-i}}{(y-i)!} \binom{Y_{t-1}}{i} a_1^i (1-a_1)^{Y_{t-1}-i}, \quad t = 1, 2, \ldots, n.$$

More generally, the p'th order model, abbreviated by INAR(p), is defined as

$$Y_t = \sum_{i=1}^{p} a_i \circ Y_{t-i} + \epsilon_t, \tag{44}$$

where $\{\epsilon_t\}$ is a sequence of i.i.d non-negative integer-valued random variables with mean μ and variance σ^2, and all p thinning operations are independent of each other; existence and generalizations of INAR(p) are studied by Latour (1997, 1998), whereas the unifying work based on convolution is presented by Joe (1996). A unique stationary and ergodic solution of (44) exists if

$$\sum_{i=1}^{p} a_i < 1. \tag{45}$$

Estimation for INAR(p) models is based on the same methods as those described for the INAR(1) model. However, in a recent contribution, Drost et al. (2009) consider the problem of semiparametric maximum likelihood estimation for INAR(p) models. In other words, the authors estimate both the finite dimensional parameters of the model plus the unknown cumulative distribution of the residual process and obtain efficient estimators. See also Jung and Tremayne (2006), Neal and Subba Rao (2007), Bu et al. (2008), and McCabe et al. (2011) for further results on estimation and prediction.

5.3. *Extensions of thinning operator-based models*

A generalization of binomial thinning is given by Joe (1996), who considers the following model

$$Y_t = A_t(Y_{t-1}; a) + \epsilon_t, \quad t = 1, 2, \ldots,$$

where $A_t(\cdot)$ is a random transformation, and $A_t(Y_{t-1}; a)$ and ϵ_t are independent. Based on this general thinning, a class of stationary moving average processes with margins in the class of infinitely divisible exponential dispersion models was introduced by Jørgensen and Song (1998).

Another generalization is that of the first-order conditional linear autoregressive process, abbreviated by CLAR(1),

$$m(Y_{t-1}) = a_1 Y_{t-1} + \mu,$$

where $m(Y_{t-1}) = \mathrm{E}[Y_t|Y_{t-1}]$, and a_1, μ are real numbers. The CLAR(1) class includes many of the non-Gaussian AR(1) models proposed in the literature and allows various generalizations of previous results; see Grunwald et al. (2000). Interestingly, when $|a_1| < 1$, the ACF of the CLAR(1) model is equal to a_1^h, $h = 1, 2, \ldots$, as in other first-order autoregressive processes including the branching process with immigration (37).

Non-negative integer-valued bilinear processes have been defined and studied by Doukhan et al. (2006); see also Latour and Truquet (2008). These processes are given by

$$Y_t = \sum_{i=1}^{p} a_i \circ Y_{t-i} + \sum_{j=1}^{q} c_j \epsilon_{t-j} + \sum_{k=1}^{m} \sum_{l=1}^{n} b_{lk} \circ (Y_{t-k}\epsilon_{t-l}) + \epsilon_t,$$

where all thinning operators are defined independently of each other and $\{\epsilon_t\}$ is a sequence of i.i.d. non-negative integer-valued random variables. Furthermore, Drost et al. (2008) consider some special cases of the above model. Random coefficient integer-valued autoregressive models have been proposed by Zheng et al. (2006, 2007). For instance, the random coefficient model of order 1 is given by

$$Y_t = a_{1t} \circ Y_{t-1} + \epsilon_t,$$

where now $\{a_{1t}\}$ is a sequence of independent and identically distributed random variables, independent of the noise $\{\epsilon_t\}$. Multivariate INAR type of models have been considered by Franke and Rao (1995), who also discuss stationarity conditions and the properties of the maximum likelihood estimator for the first-order multivariate model. For several new results regarding inference for multivariate models, see the thesis of Pedeli (2011).

5.4. *Renewal process models*

In a recent contribution by Cui and Lund (2009), the authors propose a new and simple model for stationary time series of integer counts. Their methodology does not resort to thinning operations. Instead they use a renewal process to generate a time-correlated sequence of Bernoulli trials. It turns out that superposition of i.i.d. such processes, yields stationary processes with binomial, Poisson, geometric, or any other discrete marginal distribution. Apparently, this new model class of non-Markov model offers parsimony and easily produces series with either short- or long-memory autocovariances. The model can be fitted with linear prediction techniques for stationary series.

6. Conclusions

In general, count time series refers to stochastic processes whose state space is a countable set. Although probabilistic properties of these processes are well understood, (e.g., Billingsley, 1961; Meyn and Tweedie, 1993), it is still not clear what conditions should be met for valid parametric modeling, especially when the time series is observed jointly with covariates or its behavior is driven by an unobserved process. The parametric framework allows for estimation, model assessment and forecasting by employing existing statistical software. These facts make a strong case in favor of this approach and advocate the point of view that a successful approach toward the resolution of the aforementioned problems is via the theory of GLM as advanced by Nelder and Wedderburn (1972) and McCullagh and Nelder (1989). In this contribution, we have surveyed the most commonly used models related to the regression of count time series. The foundation for this class of models is Poisson regression. However, models like (10) can be developed within the context of other distributional assumptions, like the negative binomial or other discrete distributions. The literature, both applied and theoretical, on this subject is growing fast. For instance, Andersson and Karlis (2010) consider methods for estimating the parameters of the first-order integer-valued autoregressive model in the presence of missing data and Monteiro et al. (2010) study the periodic integer-valued autoregressive model of order one. In the works of Fokianos and Fried (2010), the authors introduce the concept of intervention for the linear model (6) and discuss estimation of the intervention size as well as testing for its existence. Another problem of current interest is the analysis of multivariate count data, see Jung et al. (2011), who propose a dynamic factor model for the analysis of number of trades for five stocks from two industrial sectors.

Appendix

Proof of (33) and (35)

Consider the negative binomial regression model (31) and note that $\mathrm{E}[Y_t] = rd/(1 - a_1 - rb_1)$. Then assuming second-order stationarity, we obtain that

$$\sigma^2 = \mathrm{Var}[\epsilon_t] = r\mathrm{E}[\lambda_t + \lambda_t^2].$$

But $\mathrm{E}[\lambda_t] = \mathrm{E}[Y_t]/r$. Hence, we need to calculate $\mathrm{E}[\lambda_t^2] \equiv \mu_\lambda^{(2)}$. Consider the state equation of (31) to obtain

$$\begin{aligned}
\mu_\lambda^{(2)} &= \mathrm{E}\Big(d + a_1\lambda_{t-1} + b_1 Y_{t-1}\Big)^2 \\
&= \mathrm{E}\Big[d + (a_1 + rb_1)\lambda_{t-1} + b_1(Y_{t-1} - r\lambda_{t-1})\Big)^2 \\
&= d^2 + (a_1 + rb_1)^2\mu_\lambda^{(2)} + b_1^2\sigma^2 + 2d(a_1 + rb_1)\mathrm{E}[\lambda_t] \\
&= \Big((a_1 + rb_1)^2 + rb_1^2\Big)\mu_\lambda^{(2)} + d^2 + \Big(rb_1^2 + 2d(a_1 + rb_1)\Big)\mathrm{E}[\lambda_t].
\end{aligned}$$

Therefore, we obtain that

$$\mu_{\lambda}^{(2)} = \frac{d^2 + \Big(rb_1^2 + 2d(a_1 + rb_1)\Big)\mathrm{E}[\lambda_t]}{1 - (a_1 + rb_1)^2 - rb_1^2}.$$

Plugging this expression into the definition of σ^2 and using the fact that $d = (1 - a_1 - rb_1)\mathrm{E}[\lambda_t]$ yields to (33). Formula (35) is proved by employing representation (34) and well-known results about the ACF.

Acknowledgments

Part of this work was completed while K. Fokianos was visiting the Department of Mathematics, University of Cergy-Pontoise. The hospitality of P. Doukhan and G. Lang is greatly appreciated. Special thanks to V. Christou, M. Neumann, B. Kedem, S. Kitromilidou, and an anonymous reviewer for several suggestions that improved the presentation.

References

Agresti, A., 2002. Categorical Data Analysis, second ed. John Wiley & Sons, New York.

Al-Osh, M.A., Alzaid, A.A., 1987. First-order integer-valued autoregressive (INAR(1)) process. J. Time Ser. Anal. 8, 261–275.

Alzaid, A.A., Al-Osh, M., 1990. An integer-valued pth-order autoregressive structure (INAR(p)) process. J. Appl. Probab. 27, 314–324.

Andersson, J., Karlis, D., 2010. Treating missing values in INAR(1) models: an application to syndromic surveillance data. J. Time Ser. Anal. 31, 12–19.

Benjamin, M.A., Rigby, R.A., Stasinopoulos, D.M., 2003. Generalized autoregressive moving average models. J. Am. Stat. Assoc. 98, 214–223.

Berglund, E., Brännäs, K., 2001. Plant's entry and exit in Swedish municipalities. Ann. Reg. Sci. 35, 431–448.

Berkes, I., Horváth, L., Kokoszka, P., 2003. GARCH processes: structure and estimation. Bernoulli 9, 201–227.

Billingsley, P., 1961. Statistical Inference for Markov Processes. Univ. Chicago Press, Chicago.

Böckenholt, U., 1999. Analyzing multiple emotions over time by autoregressive negative multinomial regression models. J. Am. Stat. Assoc. 94, 757–765.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econom. 31, 307–327.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Data Analysis and Theory, second ed. Springer, New York.

Brumback, B.A., Ryan, L.M., Schwartz, J.D., Neas, L.M., Stark, P.C., Burge, H.A., 2000. Transitional regression models with application to environmental time series. J. Am. Stat Assoc. 85, 16–27.

Bu, R., McCabe, B., Hadri, K., 2008. Maximum likelihood estimation of higher–order integer-valued autoregressive process. J. Time Ser. Anal. 6, 973–994.

Cappé, O., Moulines, E., Rydén, T., 2005. Inference in Hidden Markov Models. Springer, New York.

Chandrasekhar, S., 1943. Stochastic problems in physics and astronomy. Rev. Mod. Phys. 15, 1–89.

Cox, D.R., 1981. Statistical analysis of time series: some recent developments. Scand. J. Stat. 8, 93–115.

Creal, D., Koopman, S.J., Lucas, A., 2008. A general framework for observation driven time-varying parameter models. Technical Report TI 2008–108/4, Tinbergen Institute.

Cui, Y., Lund, R., 2009. A new look at time series of counts. Biometrika 96, 781–792.

Davis, R., Wu, R., 2009. A negative binomial model for time series of counts. Biometrika 96, 735–749.

Davis, R.A., Dunsmuir, W.T.M., Streett, S.B., 2003. Observation-driven models for Poisson counts. Biometrika 90, 777–790.

Davis, R.A., Dunsmuir, W.T.M., Streett, S.B., 2005. Maximum likelihood estimation for an observation driven model for Poisson counts. Methodol. Comput. Appl. Probab. 7, 149–159.

Davis, R.A., Dunsmuir, W.T.M., Wang, Y., 2000. On autocorrelation in a Poisson regression model. Biometrika 87, 491–505.

Davis, R.A., Wang, Y., Dunsmuir, W.T.M., 1999. Modelling time series of count data. In: Ghosh, S. (Ed.), Asymptotics, Nonparametric & Time Series. Marcel Dekker, New York, pp. 63–304.

Dedecker, J., Doukhan, P., Lang, G., León, R.J.R., Louhichi, S., Prieur, C., 2007. Weak dependence: with examples and applications, Volume 190 of Lecture Notes in Statistics. Springer, New York.

Doukhan, P., Louhichi, S., 1999. A new weak dependence condition and applications to moment inequalities. Stoch. Processes Appl. 84, 313–342.

Doukhan, P., Latour, A., Oraichi, D., 2006. A simple integer-valued bilinear time series model. Adv. Appl. Probab. 38, 559–578.

Drost, F. C., van den Akker, R., Werker, B.J.M., 2008. A note on integer-valued bilinear time series models. Stat. Probab. Lett. 78, 992–996.

Drost, F.C., van den Akker, R., Werker, B.J.M., 2009. Efficient estimation of autoregression parameters and innovation distributions for semiparametric integer-valued AR(p) models. J. R. Stat. Soc. Series B 71, 467–485.

Du, J.G., Li, Y., 1991. The integer-valued autoregressive INAR(p) model. J. Time Ser. Anal. 12, 129–142.

Durbin, J., Koopman, S.J., 2001. Time Series Analysis by State Space Methods. Oxford University Press, Oxford.

Efron, B., 1986. Double exponential families and their use in generalized linear regression. J. Am. Stat. Assoc. 81, 709–721.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50, 987–1007.

Fan, J., Yao, Q., 2003. Nonlinear Time Series. Springer-Verlag, New York.

Ferland, R., Latour, A., Oraichi, D., 2006. Integer-valued GARCH processes. J. Time Ser. Anal. 27, 923–942.

Fokianos, K., 2001. Truncated Poisson regression for time series of counts. Scand. J. Stat. 28, 645–659.

Fokianos, K., Fried, R., 2010. Intereventions in INGARCH processess. J. Time Ser. Anal. 31, 210–225.

Fokianos, K., Kedem, B., 2004. Partial likelihood inference for time series following generalized linear models. J. Time Ser. Anal, 25, 173–197.

Fokianos, K., Tjøstheim, D., 2012. Nonlinear Poisson autoregression. To appear in Ann. Inst. Stat. Math.

Fokianos, K., Tjøstheim, D., 2011. Log–linear Poisson autoregression. J. Multivar. Anal. 102, 563–578.

Fokianos, K., Rahbek, A., Tjøstheim, D., 2009. Poisson autoregression. J. Am. Stat. Assoc. 104, 1430–1439.

Franke, J., 2010. Weak dependence of functional INGARCH processes. unpublished manuscript.

Franke, J., Rao, T.S., 1995. Multivariate first-order integer values autoregressions. Technical report, Department of Mathematics, UMIST.

Franke, J., Seligmann, T., 1993. Conditional maximum likelihood estimates for (INAR(1)) processes and their application to modeling epileptic seizure counts. In: Rao, T.S. (Ed.), Developments in Time Series Analysis. Chapman & Hall, London, pp. 310–330.

Grunwald, G.K., Hyndman, R.J., Tedesco, L., Tweedie, R.L., 2000. Non-Gaussian conditional linear AR(1) models. Aust. N. Z. J. Stat. 42, 479–495.

Haggan, V., Ozaki, T., 1981. Modelling nonlinear random vibrations using an amplitude-dependent autoregressive time series model. Biometrika 68, 189–196.

Harvey, A.C., Fernandes, C., 1989. Time series models for count or qualitative observations. J. Bus. Econ. Stat. 7, 407–422. with discussion.

Heinen, A., 2003. Modelling time series count data: An autoregressive conditional poisson model. Technical Report MPRA Paper 8113, University Library of Munich, Germany. available at http://mpra.ub.uni-muenchen.de/8113/.

Heyde, C.C., 1997. Quasi-Likelihood and its Applications: A General Approach to Optimal Parameter Estimation. Springer, New York.

Ispány, M., Pap, G., van Zuijlen, M.C.A., 2003. Asymptotic inference for nearly unstable INAR(1) models. J. Appl. Probab. 40, 750–765.

Joe, H., 1996. Time series models with univariate margins in the convolution–closed infinitely divisible class. J. Appl. Probab. 33, 664–677.
Jørgensen, B., Song, P.X., 1998. Stationary time series models with exponential dispersion model margins. Appl. Probab. 35, 78–92.
Jørgensen, B., Lundbye-Christensen, S., Song, P. X.-K., Sun, L., 1999. A state space model for multivariate longitudinal count data. Biometrika 86, 169–181.
Jung, R., Liesenfeld, R., Jean-François, R., 2011. Dynamic factor models for multivariate count data: an application to stock–market trading activity. J. Bus. Econ. Stat. 29, 73–85.
Jung, R.C., Tremayne, A.R., 2006. Coherent forecasting in integer time series models. Int. J. Forecast. 22, 223–238.
Jung, R.C., Tremayne, A.R., 2011. Useful models for time series of counts or simply wrong ones? AStA Adv. Stat. Anal. 95, 59–91.
Jung, R.C., Kukuk, M., Liesenfeld, R., 2006. Time series of count data: modeling, estimation and diagnostics. Comput. Stat. Data Anal. 51, 2350–2364.
Kedem, B., Chiu, L.S., 1987. On the lognormality of rain rate. Proc. Natl. Acad. Sci. U.S.A. 84, 901–905.
Kedem, B., Fokianos, K., 2002. Regression Models for Time Series Analysis. Wiley, Hoboken, NJ.
Klimko, L.A., Nelson, P.I., 1978. On conditional least squares estimation for stochastic processes. Ann. Stat. 6, 629–642.
Lambert, D., 1992. Zero–inflated Poisson regression with an application to defects is manufacturing. Technometrics 34, 1–14.
Latour, A., 1997. The multivariate GINAR(p) process. Adv. Appl. Probab. 29, 228–248.
Latour, A., 1998. Existence and stochastic structure of a non-negative integer-valued autoregressive process. J. Time Ser. Anal. 19, 439–455.
Latour, A., Truquet, L., 2008. An integer-valued bilinear type model. available at http://hal.archives-ouvertes.fr/hal-00373409/fr/.
Li, W.K., 1994. Time series models based on generalized linear models: some further results. Biometrics 50, 506–511.
MacDonald, I.L., Zucchini, W., 1997. Hidden Markov and Other Models for Discrete-valued Time Series. Chapman & Hall, London.
McCabe, B.P., Martin, G.M., Harris, D., 2011. Efficient probabilistic forecasts for counts. J. R. Stat. Soc. Ser. B 73, 253–272.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman & Hall, London.
McKenzie, E., 1985. Some simple models for discrete variate time series. Water Resour. Bull. 21, 645–650.
McKenzie, E., 1986. Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. Adv. Appl. Probab. 18, 679–705.
McKenzie, E., 1988. Some ARMA models for dependent sequences of Poisson counts. Adv. Appl. Probab. 20, 822–835.
McKenzie, E., 2003. Discrete variate time series. In: Stochastic Processes: Modelling and Simulation, Volume 21 of Handbook of Statistics. Amsterdam, North-Holland, pp. 573–606.
Meyn, S.P., Tweedie, R.L., 1993. Markov Chains and Stochastic Stability. Springer, London.
Monteiro, M., Scotto, M.G., Pereira, I., 2010. Integer-valued autoregressive processes with periodic structure. J. Stat. Plan. Inference 140, 1529–1541.
Neal, P., Subba Rao, T., 2007. MCMC for integer-valued ARMA processes. J. Time Ser. Anal. 28, 92–100.
Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. J. R. Stat. Soc. Ser. A 135, 370–384.
Neumann, M., 2011. Absolute regularity and ergodicity of Poisson count processes. Bernoulli, 17, 1268–1284.
Pedeli, X., 2011. Modelling multivariate time series for count data. Ph. D. thesis, Athens University of Economics and Business, Greece.
Priestley, M.B., 1981. Spectral Analysis and Time Series. Academic Press, London.
Rydberg, T.H., Shephard, N., 2000. A modeling framework for the prices and times of trades on the New York stock exchange. In: Fitzgerlad, W.J., Smith, R.L., Walden, A.T., Young, P.C. (Eds.), Nonlinear and Nonstationary Signal Processing. Isaac Newton Institute and Cambridge University Press, Cambridge, pp. 217–246.
Shumway, R.H., Stoffer, D.S., 2006. Time Series Analysis and its Applications, second ed. Springer, New York. With R examples.

Steutel, F.W., van Harn, K., 1979. Discrete analogues of self-decomposability and stability. Ann. Probab. 7, 893–899.

Streett, S., 2000. Some observation driven models for time series of counts. Ph. D. thesis, Colorado State University, Department of Statistics.

Teräsvirta, T., Tjøstheim, D., Granger, C.W.J., 2010. Modelling Nonlinear Economic Time Series. Oxford University Press, Oxford.

Tong, H., 1990. Nonlinear Time Series: A Dynamical System Approach. Oxford University Press, New York.

Wei, C.Z., Winnicki, J., 1990. Estimation of the means in the branching process with immigration. Ann. Stat. 18, 1757–1773.

Weiß, C.H., 2008. Thinning operations for modeling time series of counts–a survey. AStA Adv. Stat. Anal. 92, 319–341.

Weiß, C.H., 2010. INARCH(1) process: Higher-order moments and jumps. Stat. Probab. Lett. 80, 1771–1780.

West, M., Harrison, P., 1997. Bayesian Forecasting and Dynamic Models, second ed. Springer, New York.

Winnicki, J., 1986. A useful estimation theory for the branching process with immigration. Ph. D. thesis, University of Maryland, College Park, MD, USA.

Wong, W.H., 1986. Theory of partial likelihood. Ann. Stat. 14, 88–123.

Woodard, D.W., Matteson, D.S., Henderson, S.G., 2011. Stationarity of count-valued and nonlinear time series models. Electron. J. Stat. 5, 800–828.

Zeger, S.L., 1988. A regression model for time series of counts. Biometrika 75, 621–629.

Zeger, S.L., Qaqish, B., 1988. Markov regression models for time series: a quasi-likelihood approach. Biometrics 44, 1019–1031.

Zheng, H., Basawa, I.V., Datta, S., 2006. Inference for the pth-order random coefficient integer-valued process. J. Time Ser. Anal. 27, 411–440.

Zheng, H., Basawa, I.V., Datta, S., 2007. First-order random coefficient integer-valued autoregressive processes. J. Stat. Plan. Inference 137, 212–229.

Zhu, F., 2011. A negative binomial integer-valued GARCH model. J. Time Ser. Anal. 32, 54–67.

Zhu, R., Joe, H., 2006. Modelling count data time series with Markov processes based on binomial thinning. J. Time Ser. Anal. 27, 725–738.

Zhu, F., Wang, D., 2010. Diagnostic checking integer-valued ARCH(p) models using conditional residual autocorrelations. Comput. Stat. Data Anal. 54, 496–508.

This page intentionally left blank

Part VI: Nonstationary Time Series

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00013-2

Locally Stationary Processes

Rainer Dahlhaus

Institut für Angewandte Mathematik, Universität Heidelberg,
Im Neuenheimer Feld 294, 69120 Heidelberg, Germany

Abstract

The article contains an overview over locally stationary processes. At the beginning, time varying autoregressive processes are discussed in detail – both as a deep example and an important class of locally stationary processes. In the next section, a general framework for time series with time varying finite dimensional parameters is discussed with special emphasis on nonlinear locally stationary processes. Then, the paper focuses on linear processes where a more general theory is possible. First, a general definition for linear processes is given and time varying spectral densities are discussed in detail. Then, the Gaussian likelihood theory is presented for locally stationary processes. In the next section, the relevance of empirical spectral processes for locally stationary time series is discussed. Empirical spectral processes play a major role in proving theoretical results and provide a deeper understanding of many techniques. The article concludes with an overview of other results for locally stationary processes.

Keywords: locally stationary process, time varying parameter, local likelihood, derivative process, time varying autoregressive process, shape curve, empirical spectral process, time varying spectral density.

1. Introduction

Stationarity has played a major role in time series analysis for several decades. For stationary processes, there exist a large variety of models and powerful methods, such as bootstrap methods or methods based on the spectral density. Furthermore, there are important mathematical tools such as the ergodic theorem or several central limit theorems. As an example, we mention the likelihood theory for Gaussian processes which is well developed.

During recent years, the focus has turned to nonstationary time series. Here, the situation is more difficult: First, there exists no natural generalization from stationary to nonstationary time series, and second, it is often not clear how to set down a meaningful asymptotics for nonstationary processes. An exception are nonstationary models that are generated by a time invariant generation mechanism – for example, integrated or cointegrated models. These models have attracted a lot of attention during recent years. For general, nonstationary processes ordinary asymptotic considerations are often contradictory to the idea of nonstationarity since future observations of a nonstationary process may not contain any information at all on the probabilistic structure of the process at present. For this reason, the theory of locally stationary processes is based on infill asymptotics originating from nonparametric statistics.

As a consequence valuable asymptotic concepts such as consistency, asymptotic normality, efficiency, LAN expansions, neglecting higher-order terms in Taylor expansions, etc. can be used in the theoretical treatment of statistical procedures for such processes. This leads to several meaningful results also for the original nonrescaled case such as the comparison of different estimates, the approximations for the distribution of estimates and bandwidth selection (for a detailed example, see Remark 2).

The type of processes that can be described with this infill asymptotics are processes which locally at each time point are close to a stationary process but whose characteristics (covariances, parameters, etc.) are gradually changing in an unspecific way as time evolves. The simplest example for such a process may be an AR(p) process whose parameters are varying in time. The infill asymptotic approach means that time is rescaled to the unit interval. For time varying AR processes, this is explained in detail in the next section. Another example are GARCH processes that have recently been investigated by several authors – see Section 3.

The idea of having locally approximately a stationary process was also the starting point of Priestley's (1965) theory of processes with evolutionary spectra (see also Priestley (1988), Granger and Hatanaka (1964), Tjøstheim (1976), and Mélard and Herteleer-de-Schutter (1989), among others). Priestley considered processes having a time varying spectral representation

$$X_t = \int_{-\pi}^{\pi} \exp(i\lambda t)\, \tilde{A}_t(\lambda)\, d\xi(\lambda), \quad t \in \mathbf{Z}$$

with an orthogonal increment process $\xi(\lambda)$ and a time varying transfer function $\tilde{A}_t(\lambda)$. (Priestley mainly looked at continuous-time processes, but the theory is the same). Also within this approach, asymptotic considerations (e.g., for judging the efficiency of a local covariance estimator) are not possible or meaningless from an applied view. Using the above mentioned infill asymptotics means, in this case, basically to replace $\tilde{A}_t(\lambda)$ with some function $A(t/T, \lambda)$ – see (78).

Beyond the above cited references on processes with evolutionary spectra, there has also been work on processes with time varying parameters which does not use the infill asymptotics discussed in this paper (cf. Subba Rao (1970) and Hallin (1986), among others). Furthermore, there have been several papers on inference for processes with time varying parameters – mainly within the engineering literature (cf. Grenier (1983) and Kayhan et al. (1994), among others).

The paper is organized as follows: In Section 2, we start with time varying autoregressive processes as a deep example and an important class of locally stationary processes. There we mark many principles and problems addressed at later stages with higher generality. In Section 3, we present a more general framework for time series with time varying finite-dimensional parameters and show how nonparametric inference can be done and theoretically handled. We also introduce derivative processes that play a major role in the derivations. The results cover in particular nonlinear processes such as GARCH processes with time varying parameters.

If one restrict to linear processes or even more to Gaussian processes, then a much more general theory is possible which is developed in the subsequent sections. In Section 4, we give a general definition for linear processes and discuss time varying spectral densities in detail. Section 5 then contains the Gaussian likelihood theory for locally stationary processes. In Section 6, we discuss the relevance of empirical spectral processes for locally stationary time series. Empirical spectral processes play a major role in proving theoretical results and provide a deeper understanding of many techniques.

2. Time varying autoregressive processes – A deep example

We now discuss time varying autoregressive processes in detail. In particular, we mark many principles and problems addressed at later stages with higher generality. Consider the time varying AR(1) process

$$X_t + \alpha_t X_{t-1} = \sigma_t \varepsilon_t \quad \text{with } \varepsilon_t \text{ i.i.d. } \mathcal{N}(0,1). \tag{1}$$

We now apply infill asymptotics, that is, we rescale the parameter curves α_t and σ_t to the unit interval. This means that we replace them by $\alpha(\frac{t}{T})$ and $\sigma(\frac{t}{T})$ with curves $\alpha(\cdot)\colon [0,1] \to (-1,1)$ and $\sigma(\cdot)\colon [0,1] \to (0,\infty)$ leading in the general AR(p) case to the definition given in (2) below. Formally, this results in replacing X_t by a triangular array of observations $\big(X_{t,T}\,;\ t = 1,\ldots,T;\ T \in \mathbb{N}\big)$, where T is the sample size.

We now indicate again the reason for this rescaling. Suppose we fit the parametric model $\alpha_{\theta,t} := b + ct + dt^2$ to the nonrescaled model (1), which we assume to be observed for $t = 1,\ldots,T$. It is easy to construct different estimators for the parameters (e.g., the least squares estimator, the maximum likelihood estimator or a moment estimator), but it is nearly impossible to derive the finite sample properties of these estimators. On the other hand, classical nonrescaled asymptotic considerations for comparing these estimators make no sense since with $t \to \infty$ also $\alpha_{\theta,t} \to \infty$, while, e.g., $|\alpha_t|$ may be less than one within the observed segment – i.e., the resulting asymptotic results are without any relevance for the observed stretch of data. By rescaling α_t and σ_t to the unit interval as described above, we overcome these problems. As T tends to infinity, more and more observations of each local structure become available, and we obtain a reasonable framework for a meaningful asymptotic analysis of statistical procedures allowing to retain such powerful tools as consistency, asymptotic normality, efficiency, LAN expansions, etc. for nonstationary processes. For example, the results on asymptotic normality of an estimator obtained in this framework may

be used to approximate the distribution of the estimator in the finite sample situation. It is important to note that classical asymptotics for stationary processes arises as a special case of this infill asymptotics in case where all parameter curves are constant.

Unfortunately, infill asymptotics does not describe the physical behavior of the process as $T \to \infty$. This may be unusual for time series analysis, but it has been common in other branches of statistics for many years. We remark that all statistical methods and procedures stay the same or can easily be translated from the rescaled processes to the original nonrescaled processes. A more complicated example on how the results of the rescaled case transfer to the nonrescaled case is given in Remark 2.

In the following, we, therefore, consider time varying autoregressive $\big(\text{tvAR(p)}\big)$ processes defined by

$$X_{t,T} + \sum_{j=1}^{p} \alpha_j \left(\frac{t}{T}\right) X_{t-j,T} = \sigma\left(\frac{t}{T}\right) \varepsilon_t, \quad t \in \mathbf{Z}, \tag{2}$$

where the ε_t are independent random variables with mean zero and variance 1. We assume $\sigma(u) = \sigma(0)$, $\alpha_j(u) = \alpha_j(0)$ for $u < 0$ and $\sigma(u) = \sigma(1)$, $\alpha_j(u) = \alpha_j(1)$ for $u > 1$. In addition, we usually assume some smoothness conditions on $\sigma(\cdot)$ and the $\alpha_j(\cdot)$. In addition, one may include a time varying mean by replacing $X_{t-j,T}$ in (2) by $X_{t-j,T} - \mu(t - j/T)$ – see Section 7.6.

In some neighborhood of a fixed time point $u_0 = t_0/n$, the process $X_{t,T}$ can be approximated by the stationary process $\tilde{X}_t(u_0)$ defined by

$$\tilde{X}_t(u_0) + \sum_{j=1}^{p} \alpha_j(u_0)\, \tilde{X}_{t-j}(u_0) = \sigma(u_0)\, \varepsilon_t, \quad t \in \mathbf{Z}. \tag{3}$$

It can be shown (see Section 3) that we have under suitable regularity conditions

$$\left|X_{t,T} - \tilde{X}_t(u_0)\right| = O_p\Big(\Big|\frac{t}{T} - u_0\Big| + \frac{1}{T}\Big) \tag{4}$$

which justifies the notation "locally stationary process." $X_{t,T}$ has an unique time varying spectral density which is locally the same as the spectral density of $\tilde{X}_t(u)$, namely

$$f(u,\lambda) := \frac{\sigma^2(u)}{2\pi} \left|1 + \sum_{j=1}^{p} \alpha_j(u) \exp(-ij\lambda)\right|^{-2} \tag{5}$$

(see Example 7). Furthermore, it has locally in some sense the same autocovariance

$$c(u,j) := \int_{-\pi}^{\pi} e^{ij\lambda} f(u,\lambda) d\lambda, \quad j \in \mathbf{Z}$$

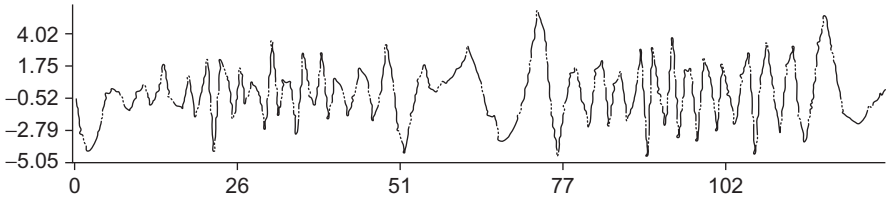

Fig. 1. $T = 128$ as realizations of a time varying AR(2) model.

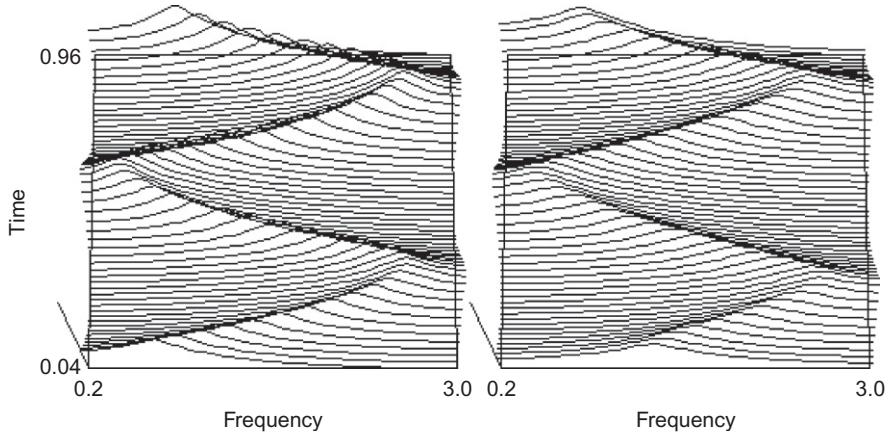

Fig. 2. True and estimated time varying spectrum of a tvAR(2) process.

since $\operatorname{cov}(X_{[uT],T}, X_{[uT]+k,T}) = c(u,k) + O(T^{-1})$ uniformly in u and k (cf. (73)). This justifies to term $c(u,k)$, the local covariance function of $X_{t,T}$ at time $u = t/T$.

As an example, Fig. 1 shows $T = 128$ observations of a tvAR(2) process with mean 0 and parameters $\sigma(u) \equiv 1$, $\alpha_1(u) \equiv -1.8\cos(1.5 - \cos 4\pi u)$, $\alpha_2(u) = 0.81$, and Gaussian innovations ε_t. The parameters are chosen in a way such that for fixed u the complex roots of the characteristic polynomial are $\frac{1}{0.9}\exp[\pm i(1.5 - \cos 4\pi u)]$, that is, they are close to the unit circle and their phase varies cyclically with u. As could be expected from these roots, the observations show a periodic behavior with time varying period length. The left picture of Fig. 2 shows the true time varying spectrum of the process. One clearly sees that the location of the peak is also time varying (it is located at frequency $1.5 - \cos 4\pi u$).

2.1. Local estimation by stationary methods on segments

An ad-hoc method that works in nearly all cases for locally stationary processes is to do inference via stationary methods on segments. The idea is that the process $X_{t,T}$ is almost stationary on a reasonably small segment $\{t\colon |t/T - u_0| \le b/2\}$. The parameter of interest (or the correlation, spectral density, etc.) is estimated by some classical

method, and the resulting estimate is assigned to the midpoint u_0 of the segment. By shifting the segment, this finally leads to an estimate of the unknown parameter curve (time varying correlation, time varying spectral density, etc). An important modification of this method is obtained when more weight is put on data in the center of the interval than at the edges. This can often be achieved by using a data taper on the segment or by using a kernel-type estimate.

Since we use observations from the process $X_{t,T}$ (instead of $\tilde{X}_t(u_0)$), the procedure causes a bias which depends on the degree of nonstationarity of the process on the segment. It is possible to evaluate this bias and to use the resulting expression for an optimal choice of the segment length. To demonstrate this, we now discuss the estimation of the AR coefficient functions by classical Yule-Walker estimates on segments. Since the approximating process $\tilde{X}_t(u_0)$ is stationary, we obtain from (3) that the Yule-Walker equations hold locally at time u_0, that is, we have with $\boldsymbol{\alpha}(u_0) := \big(\alpha_1(u_0), \ldots, \alpha_p(u_0)\big)'$

$$\boldsymbol{\alpha}(u_0) = -R(u_0)^{-1} r(u_0) \quad \text{and} \quad \sigma^2(u_0) = c(u_0, 0) + \boldsymbol{\alpha}(u_0)' r(u_0), \tag{6}$$

where $r(u_0) := \big(c(u_0, 1), \ldots, c(u_0, p)\big)'$ and $R(u_0) := \{c(u_0, i-j)\}_{i,j=1,\ldots,p}$.

To estimate $\boldsymbol{\alpha}(u_0)$, we use the classical Yule-Walker estimator on the segment $[u_0T] - N/2 + 1, \ldots, [u_0T] + N/2$ (ordinary time) or on $[u_0 - b_T/2, u_0 + b_T/2]$ (rescaled time with bandwidth $b_T := N/T$), that is

$$\hat{\boldsymbol{\alpha}}_T(u_0) = -\hat{R}_T(u_0)^{-1} \hat{r}_T(u_0) \quad \text{and} \quad \hat{\sigma}_T^2(u_0) = \hat{c}_T(u_0, 0) + \hat{\boldsymbol{\alpha}}_T(u_0)' \hat{r}_T(u_0), \tag{7}$$

where $\hat{r}_T(u_0) := (\hat{c}_T(u_0, 1), \ldots, \hat{c}_T(u_0, p))'$ and $\hat{R}_T(u_0) := \{\hat{c}_T(u_0, i-j)\}_{i,j=1,\ldots,p}$ with some covariance estimator $\hat{c}_T(u_0, j)$.

Before we discuss the properties of this estimator, we first discuss different covariance estimates and their properties.

2.2. *Local covariance estimation*

The covariance estimate with data taper on the segment $[u_0T] - N/2 + 1, \ldots, [u_0T] + N/2$ is

$$\hat{c}_T(u_0, k) := \frac{1}{H_N} \sum_{\substack{s,t=1 \\ s-t=j}}^{N} h\left(\frac{s}{N}\right) h\left(\frac{t}{N}\right) X_{[u_0T]-\frac{N}{2}+s,T} \, X_{[u_0T]-\frac{N}{2}+t,T}. \tag{8}$$

where $h : [0, 1] \to \mathbf{R}$ is a data taper with $h(x) = h(1 - x)$, $H_N := \sum_{j=0}^{N-1} h^2(\frac{j}{N}) \sim N \int_0^1 h^2(x)\, dx$ is the normalizing factor. The data taper usually is largest at $x = 1/2$ and decays slowly to 0 at the edges. For $h(x) = \chi_{(0,1]}(x)$, we obtain the classical nontapered covariance estimate.

An asymptotically equivalent (and from a certain viewpoint more intuitive estimator) is the kernel density estimator

$$\tilde{c}_T(u_0, k) := \frac{1}{b_T T} \sum_t K\left(\frac{u_0 - (t + k/2)/T}{b_T}\right) X_{t,T} X_{t+k,T} \tag{9}$$

where $K: \mathbf{R} \to [0, \infty)$ is a kernel with $K(x) = K(-x)$, $\int K(x)dx = 1$, $K(x) = 0$ for $x \notin [-1/2, 1/2]$ and b_T is the bandwidth. Also, equivalent is

$$\tilde{\tilde{c}}_T(u_0, i, j) := \frac{1}{b_T T} \sum_t K\left(\frac{u_0 - t/T}{b_T}\right) X_{t-i,T} X_{t-j,T} \tag{10}$$

with $i - j = k$, which appears in least square regression – cf. Example 1(i). If $K(x) = h(x)^2$, all three estimators are equivalent in the sense that they lead to the same asymptotic bias, variance, and mean-squared error. For reasons of clarity, a few remarks are in order:

1. The classical stationary method on a segment is in this case the estimator without data taper which is the same as the kernel estimator with a rectangular kernel.
2. A first step toward a better estimate (as it is proved below) is to put higher weights in the middle and lower weights at the edges of the observation domain in order to cope in a better way with the nonstationarity of $X_{t,T}$ on the segment. In this context, this may be either achieved by using a kernel estimate or a data taper, which is asymptotically equivalent. This is straightforward for local covariance estimates and local Yule-Walker estimates and can usually also be applied to other estimation problems.
3. Data tapers have also been used for stationary time series (in particular in spectral estimation, but also with Yule-Walker estimates and covariance estimation where they give positive definite autocovariances with a lower bias). Thus, the reason for using data tapers for segment estimates is fold: reducing the bias due to nonstationarity on the segment and reducing the (classical) bias of the procedure as a stationary method.

We now determine the mean-squared error of the above estimators. Furthermore, we determine the optimal segment length N and show that weighted estimates are better than ordinary estimates.

THEOREM 1. *Suppose $X_{t,T}$ is locally stationary with mean 0. Under suitable regularity conditions (in particular second-order smoothness of $c(\cdot, k)$), we have for $\hat{c}_T(u_0, k)$, $\tilde{c}_T(u_0, k)$, and $\tilde{\tilde{c}}_T(u_0, i, j)$ with $K(x) = h(x)^2$ and $b_T = N/T$*

$$(i) \quad \mathbf{E}\hat{c}_T(u_0, k) = c(u_0, k) + \frac{1}{2} b_T^2 \int x^2 K(x) dx \left[\frac{\partial^2}{\partial^2 u} c(u_0, k)\right] + o(b_T^2) + O\left(\frac{1}{b_T T}\right)$$

and

$$(ii)\quad \operatorname{var}\big(\hat{c}_T(u_0,k)\big) = \frac{1}{b_T T}\int_{-1/2}^{1/2} K(x)^2 dx \sum_{\ell=-\infty}^{\infty} c(u_0,\ell)\big[c(u_0,\ell)+c(u_0,\ell+2k)\big] + o\left(\frac{1}{b_T T}\right).$$

Proof. (i) see Dahlhaus (1996c), (ii) is omitted (the form of the asymptotic variance is the same as in the stationary case). □

Note that the above bias of order b_T^2 is solely due to nonstationarity, which is measured by $\partial^2/\partial u^2 c(u_0,k)$. If the process is stationary, this second derivative is zero and the bias disappears. The bandwidth b_T may now be chosen to minimize the mean-squared error.

Remark 1 (Minimizing the mean-squared error). Let $\mu(u_0) := \frac{\partial^2}{\partial^2 u_0} c(u_0,k)$, $\tau(u_0) := \sum_{\ell=-\infty}^{\infty} c(u_0,\ell)\big[c(u_0,\ell)+c(u_0,\ell+2k)\big]$, $d_K := \int x^2 K(x)\,dx$ and $v_K := \int K(x)^2\,dx$. Then, we have for the mean-squared error

$$\mathbf{E}\left|\hat{c}_T(u_0,k)-c(u_0,k)\right|^2 = \frac{b^4}{4} d_K^2\,\mu(u_0)^2 + \frac{1}{bT} v_K\,\tau(u_0) + o\left(b^4+\frac{1}{bT}\right). \tag{11}$$

It can be shown (cf. Priestley, 1981, Chapter 7.5) that this MSE gets minimal for

$$K(x) = K_{opt}(x) = 6x(1-x), \quad 0 \le x \le 1 \tag{12}$$

and

$$b = b_{opt}(u_0) = C(K_{opt})^{1/5}\left[\frac{\tau(u_0)}{\mu(u_0)^2}\right]^{1/5} T^{-1/5} \tag{13}$$

where $C(K) = v_K/d_K^2$. In this case, we have with $c(K) = v_K d_K^{1/2}$

$$T^{4/5}\,\mathbf{E}\left|\hat{c}_T(u_0,k)-c(u_0,k)\right|^2 = \frac{5}{4} c(K_{opt})^{4/5}\,\mu(u_0)^{2/5}\,\tau(u_0)^{4/5} + o(1). \tag{14}$$

$\mu(u_0) = \frac{\partial^2}{\partial^2 u_0} c(u_0,k)$ measures the "degree of nonstationarity," while $\tau(u_0)$ measures the variability of the estimate at time u_0. The segment length $N_{opt} = b_{opt} T$ gets larger if $\mu(u_0)$ gets smaller, i.e., if the process is closer to stationarity (in this case: if the kth

order covariance is more constant/more linear in time). At the same time, the mean-squared error decreases. The results are similar to kernel estimation in nonparametric regression. A yet unsolved problem is how to adaptively determine the bandwidth from the observed process. □

2.3. *Segment selection and asymptotic mean-squared error for local Yule-Walker estimated*

For the local Yule-Walker estimates from (7) with the covariances $\hat{c}_T(u_0, k)$ as defined in (8), Dahlhaus and Giraitis (1998) have proved (see also Example 4)

$$\mathbf{E}\,\hat{\boldsymbol{\alpha}}_T(u_0) = \boldsymbol{\alpha}(u_0) - \frac{b^2}{2}\, d_K\, \boldsymbol{\mu}(u_0) + o(b^2)$$

with

$$\boldsymbol{\mu}(u_0) = R(u_0)^{-1}\left[\left(\frac{\partial^2}{\partial u^2} R(u)\right)\boldsymbol{\alpha}(u_0) + \left(\frac{\partial^2}{\partial u^2} r(u)\right)\right]_{u=u_0}$$

and

$$\operatorname{var}\left(\hat{\boldsymbol{\alpha}}_T(u_0)\right) = \frac{1}{bT}\, v_K\, \sigma^2(u_0)\, R(u_0)^{-1} + o\left(\frac{1}{bT}\right).$$

Thus, we obtain for $\mathbf{E}\left\|\hat{\boldsymbol{\alpha}}_T(u_0) - \boldsymbol{\alpha}(u_0)\right\|^2$, the same expression as in (11) with $\tau(u_0) = \sigma^2(u_0)\,\mathrm{tr}\{R(u_0)^{-1}\}$ and $\mu(u_0)^2$ replaced by $\|\boldsymbol{\mu}(u_0)\|^2$. With these changes, the optimal bandwidth is given by (13) and the optimal mean-squared error by (14).

Remark 2 (Implications for nonrescaled processes). Suppose that we observe data from a (nonrescaled) tvAR(p) process

$$X_t + \sum_{j=1}^{p} \alpha_{tj}\, X_{t-j} = \sigma_t\, \varepsilon_t, \quad t \in \mathbf{Z}. \tag{15}$$

In order to estimate $\boldsymbol{\alpha}_t$ at some time t_0, we may use the segment Yule-Walker estimator as given in (7). The theoretically optimal segment length is given by (13) as

$$N_{opt}(u_0) = C(K_{opt})^{1/5}\left[\frac{\tau(u_0)}{\|\boldsymbol{\mu}(u_0)\|^2}\right]^{1/5} T^{4/5}, \tag{16}$$

which at first sight depends on T and the rescaling.
Suppose that we have parameter functions $\tilde{a}_j(\cdot)$ and some $T > t_0$ with $\tilde{a}_j(\frac{t_0}{T}) = \alpha_j(t_0)$ (i.e., the original function has been rescaled to the unit interval) and we denote by $\tilde{R}$,

$\tilde{r}$, and $\tilde{\boldsymbol{\alpha}}$, the corresponding parameters in the rescaled world (i.e., $\tilde{R}(u_0) = R(t_0)$ etc.). Then,

$$\tau(u_0) = \tilde{\sigma}^2(u_0)\,\mathrm{tr}\{\tilde{R}(u_0)^{-1}\} = \sigma^2(t_0)\,\mathrm{tr}\{R(t_0)^{-1}\}$$

and (with the second-order difference as an approximation of the second derivative)

$$\begin{aligned}\boldsymbol{\mu}(u_0) &= \tilde{R}(u_0)^{-1}\left[\left(\frac{\partial^2}{\partial u^2}\tilde{R}(u)\right)\tilde{\boldsymbol{\alpha}}(u_0) + \left(\frac{\partial^2}{\partial u^2}\tilde{r}(u)\right)\right]_{u=u_0}\\ &\approx R(t_0)^{-1}\left[\frac{R(t_0) - 2R(t_0-1) + R(t_0-2)}{1/T^2}\,\boldsymbol{a}(t_0)\right.\\ &\quad\left. + \frac{r(t_0) - 2r(t_0-1) + r(t_0-2)}{1/T^2}\right].\end{aligned}$$

Plugging this into (16) reveals that T drops out completely, and the optimal segment length can completely be determined in terms of the original nonrescaled process. This is a nice example on how the asymptotic considerations in the rescaled world can be transferred with benefit to the original nonrescaled world. □

These considerations justify the asymptotic approach of this paper: while it is not possible to set down a meaningful asymptotic theory for the nonrescaled model (1), an approach using the rescaled model (2) leads to meaningful results also for the model (1). Another example for this relevance is the construction of confidence intervals for the local Yule-Walker estimates from the central limit theorem by Dahlhaus and Giraitis (1998) Theorem 3.2.

2.4. *Parametric Whittle-type estimates – A first approach*

We now assume that the $p+1$-dimensional parameter curve $\boldsymbol{\theta}(\cdot) = \big(\alpha_1(\cdot), \ldots, \alpha_p(\cdot), \sigma^2(\cdot)\big)'$ is parameterized by a finite-dimensional parameter $\eta \in \mathbf{R}^q$, that is, $\boldsymbol{\theta}(\cdot) = \boldsymbol{\theta}_\eta(\cdot)$. An example studied below is where the AR coefficients are modeled by polynomials. Another example is where the AR coefficients are modeled by a parametric transition curve as in Section 2.6(iv). In particular, when the length of the time series is short, this may be a proper choice. We now show how the stationary Whittle likelihood can be generalized to the locally stationary case (another generalization is given in (89)).

If we were looking for a *nonparametric* estimate for the parameter curve $\boldsymbol{\theta}(\cdot)$, we could apply the stationary Whittle estimate on a segment leading to

$$\hat{\boldsymbol{\theta}}_T^W(u_0) := \operatorname*{argmin}_{\theta\in\Theta} \mathcal{L}_T^W(u_0, \boldsymbol{\theta}) \tag{17}$$

with the Whittle likelihood

$$\mathcal{L}_T^W(u_0, \boldsymbol{\theta}) := \frac{1}{4\pi}\int_{-\pi}^{\pi}\left\{\log 4\pi^2 f_{\boldsymbol{\theta}}(\lambda) + \frac{I_T(u_0,\lambda)}{f_{\boldsymbol{\theta}}(\lambda)}\right\} d\lambda \tag{18}$$

with the tapered periodogram on a segment about u_0, that is

$$I_T(u_0, \lambda) := \frac{1}{2\pi H_N} \left| \sum_{s=1}^{N} h\Big(\frac{s}{N}\Big) X_{[u_0 T]-N/2+s,T} \exp\big(-i\lambda s\big) \right|^2. \tag{19}$$

Here, $h(\cdot)$ is a data taper as in (8). For $h(x) = \chi_{(0,1]}(x)$, we obtain the nontapered periodogram. The properties of this nonparametric estimate are discussed later – in particular in Example 3 and at the end of Example 9. In case of a tvAR(p) process, $\hat{\boldsymbol{\theta}}_T(u_0)$ is exactly the local Yule-Walker estimate defined in (7) with the covariance estimate given in (8).

Suppose now that we want to fit globally the parametric model $\boldsymbol{\theta}(\cdot) = \boldsymbol{\theta}_\eta(\cdot)$ to the data, that is, we have the time varying spectrum $f_\eta(u, \lambda) := f_{\boldsymbol{\theta}_\eta(u)}(\lambda)$. Since $\mathcal{L}_T^W(u, \boldsymbol{\theta})$ is an approximation of the Gaussian log-likelihood on the segment $\{[uT] - N/2 + 1, \ldots, [uT] + N/2\}$, a reasonable approach is to use

$$\hat{\eta}_T^{BW} := \operatorname*{argmin}_{\eta \in \Theta_\eta} \mathcal{L}_T^{BW}(\eta) \tag{20}$$

with the *block Whittle likelihood*

$$\mathcal{L}_T^{BW}(\eta) := \frac{1}{4\pi} \frac{1}{M} \sum_{j=1}^{M} \int_{-\pi}^{\pi} \left\{ \log 4\pi^2 f_\eta(u_j, \lambda) + \frac{I_T(u_j, \lambda)}{f_\eta(u_j, \lambda)} \right\} d\lambda. \tag{21}$$

Here, $u_j := t_j/T$ with $t_j := S(j-1) + N/2 \;\; (j = 1, \ldots, M)$, i.e., we calculate the likelihood on overlapping segments which we shift each time by S. Furthermore, $T = S(M-1) + N$. A better justification of the form of the likelihood is provided by the asymptotic Kullback-Leibler information divergence derived in Theorem 5.

As discussed above, the reason for using data tapers is twofold: they reduce the bias due to nonstationarity on the segment and they reduce the leakage (already known from the stationary case). It is remarkable that the taper in this case does not lead to an increase of the asymptotic variance if the segments are overlapping (cf. Dahlhaus, 1997, Theorem 3.3).

The properties of the above estimate are discussed by Dahlhaus (1997) including consistency, asymptotic normality, model selection, and the behavior if the model is misspecified. The estimate is asymptotically efficient if $S/N \to 0$.

As an example, we now fit a tvAR(p) model to the data from Fig. 1 and estimate the parameters by minimizing $\mathcal{L}_T^{BW}(\eta)$. The AR coefficients are modeled as polynomials with different orders. Thus, we fit the model

$$\alpha_j(u) = \sum_{k=0}^{K_j} b_{jk}\, u^k \quad (j = 1, \ldots, p) \quad \text{and} \quad \sigma(u) \equiv c$$

Table 1
Values for AIC for $p = 2$ and different polynomial orders.

K_2 \ K_1	4	5	6	7	8	9
0	0.929	0.888	0.669	0.685	0.673	0.689
1	0.929	0.901	0.678	0.694	0.682	0.698
2	0.916	0.888	0.694	0.709	0.697	0.712

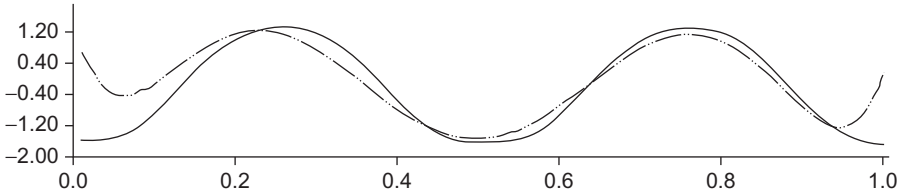

Fig. 3. True and estimated parameter curve $\alpha_1(\cdot)$.

to the data. The model orders $p, K_1, \ldots, K_p$ are chosen by minimizing the AIC criterion

$$AIC(p, K_1, \ldots, K_p) = \log \hat{\sigma}^2(p, K_1, \ldots, K_p) + 2 \left(p + 1 + \sum_{j=1}^{p} K_j \right) / T.$$

Table 1 shows these values for $p = 2$ and different K_1 and K_2. The values for other p turned out to be larger. Thus, a model with $p = 2, K_1 = 6, K_2 = 0$ is fitted. The function $\alpha_1(u)$ and its estimate are plotted in Fig. 3. For $\hat{a}_2(u)$, we obtain 0.71 (a constant is fitted because of $K_2 = 0$), while the true $\alpha_2(u)$ is 0.81. Furthermore, $\hat{\sigma}^2 = 1.71$, while $\sigma^2 = 1.0$. The corresponding (parametric) estimate of the spectrum is the right picture of Fig. 2, and the difference to the true spectrum is plotted in Fig. 4.

Given the small sample size, the quality of the fit is remarkable. Two negative effects can be observed. First, the fit of $\alpha_1(u)$ becomes rather bad outside $u_1 = 0.063$ and $u_M = 0.938$. This is not surprising, due to the behavior of a polynomial and the fact that the use of $\mathcal{L}_T^{BW}(\eta)$ as a distance only punishes bad fits inside the interval $[u_1, u_M]$. This end effect improves if one chooses $K_1 = 8$ instead of $K_1 = 6$. A better way seems to modify $\mathcal{L}_T^{BW}(\eta)$ and to include periodograms of shorter lengths at the edges. The second effect is that the peak in the spectrum is underestimated. This bias is in part due to the nonstationarity of the process on intervals $(u_j - N/(2T), u_j + N/(2T))$, where $I_T(u_j, \lambda)$ is calculated.

We mention that the above estimates can be written in closed form and calculated without an optimization routine. More generally, this holds for tvAR(p) models if σ^2 is constant and $\alpha_j(u) = \sum_{k=1}^{K} b_{jk} f_k(u)$ with some functions $f_1(u), \ldots, f_K(u)$ (in the above case, $f_k(u) = u^{k-1}$). For details, see Dahlhaus (1997), Section 4.

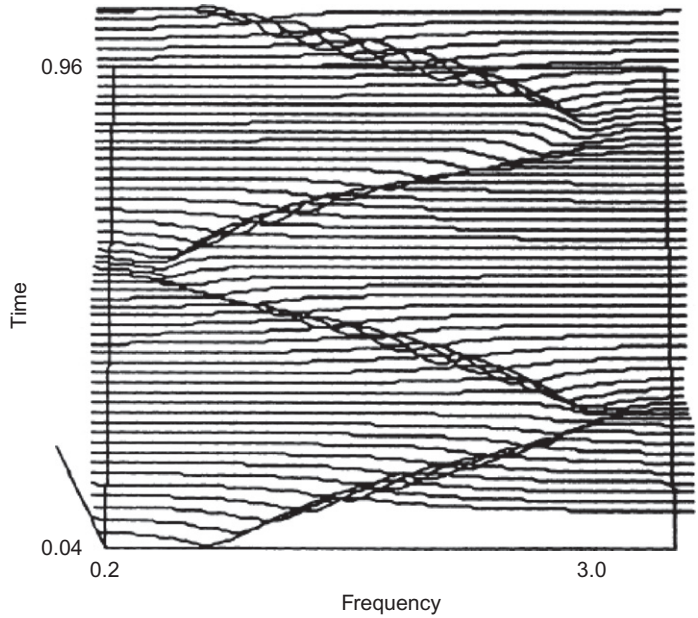

Fig. 4. Difference of estimated and true spectrum.

A closer look at the above estimate reveals that it is somehow the outcome of a two-step procedure, where in the first step, the periodogram is calculated on segments (which implicitly includes some smoothing with bandwidth $b = N/T$), and afterward the AR(p) process with the above polynomials is fitted to the outcome (instead of a direct fit of the AR(p) model and the polynomials to the data). We now make this more precise.

With the above form of the spectrum $f_\eta(u,\lambda)$ (cf. (5)) and Kolmogorov's formula, (cf. Brockwell and Davis, 1991, Theorem 5.8.1), we obtain with $\hat{R}_T(u_j)$ and $\hat{r}_T(u_j)$ as defined in (7) after some straightforward calculations

$$\begin{aligned}\mathcal{L}_T^{BW}(\eta) = \frac{1}{2}\frac{1}{M}\sum_{j=1}^{M}\Big[&\log 4\pi^2\sigma_\eta^2(u_j) + \frac{1}{\sigma_\eta^2(u_j)} \\ &\times \Big(\hat{c}_T(u_j,0) - \hat{r}_T(u_j)'\,\hat{R}_T(u_j)^{-1}\hat{r}_T(u_j)\Big)\Big] \\ &+ \frac{1}{2}\frac{1}{M}\sum_{j=1}^{M}\frac{1}{\sigma_\eta^2(u_j)}\Big[\big(\hat{R}_T(u_j)\,\boldsymbol{\alpha}_\eta(u_j) + \hat{r}_T(u_j)\big)'\,\hat{R}_T(u_j)^{-1} \\ &\times \big(\hat{R}_T(u_j)\,\boldsymbol{\alpha}_\eta(u_j) + \hat{r}_T(u_j)\big)\Big].\end{aligned}$$

We now plug in the Yule-Walker estimate $\hat{\boldsymbol{\alpha}}_T(u) = -\hat{R}_T(u)^{-1}\,\hat{r}_T(u)$ with asymptotic variance proportional to $\sigma^2(u)\,R(u)^{-1}$ and $\hat{\sigma}_T^2(u) = \hat{c}_T(u,0) - \hat{r}_T(u)'\,\hat{R}_T(u)^{-1}\,\hat{r}_T(u)$ with asymptotic variance $2\,\sigma^2(u)$. Since $\log x = (x-1) - \frac{1}{2}(x-1)^2 + o((x-1)^2)$, we obtain

$$\mathcal{L}_T^{BW}(\eta) = \frac{1}{2}\frac{1}{M}\sum_{j=1}^{M}\frac{1}{2\,\sigma_\eta^4(u_j)}\Big[\sigma_\eta^2(u_j) - \hat{\sigma}_T^2(u_j)\Big]^2$$

$$+ \frac{1}{2}\frac{1}{M}\sum_{j=1}^{M}\Big[\Big(\boldsymbol{\alpha}_\eta(u_j) - \hat{\boldsymbol{\alpha}}_T(u_j)\Big)'\sigma_\eta^2(u_j)^{-1}\,\hat{R}_T(u_j)\Big(\boldsymbol{\alpha}_\eta(u_j) - \hat{\boldsymbol{\alpha}}_T(u_j)\Big)\Big]$$

$$+ \frac{1}{2}\frac{1}{M}\sum_{j=1}^{M}\log 4\pi^2\hat{\sigma}_T^2(u_j) + \frac{1}{2} + o\left(\left(\frac{\sigma_\eta^2(u_j) - \hat{\sigma}_T^2(u)}{\sigma_\eta^2(u_j)}\right)^2\right).$$

If the model is correctly specified, then we have for η close to the minimum: $\sigma_\eta^2(u_j)^{-1}\,\hat{R}_T(u_j) \approx \sigma^2(u_j)^{-1}\,R(u_j)$ and $2\,\sigma_\eta^4(u_j) \approx 2\,\sigma^2(u_j)$, which means that $\hat{\eta}_T$ is approximately obtained by a weighted least squares fit of $\boldsymbol{\alpha}_\eta(u)$ and $\sigma_\eta^2(u)$ to the Yule-Walker estimates on the segments. The method works in this case since the (parametric!) model fitted in the second step is somehow "smoother" than the first smoothing implicitly induced by using the periodogram on a segment. However, we would clearly run into problems if the fitted polynomials were of high order or if even $K_j = K_j(T) \to \infty$ as $T \to \infty$.

A good alternative seems to use the quasi-likelihood $\mathcal{L}_T^{GW}(\eta)$ from (89) or (in particular for AR(p) models) the conditional likelihood estimate from (30) with $\ell_{t,T}(\cdot)$ as in (23) for which the estimator can explicitly be calculated if $\sigma(\cdot) \equiv c$. For $\sigma_0(\cdot) \neq c$, iterative or approximative solutions are needed. The properties of this estimator have not been investigated yet. In any case, the benefit of the likelihood $\mathcal{L}_T^{BW}(\eta)$ and even more of the improved likelihood $\mathcal{L}_T^{GW}(\eta)$ is their generality because they can be applied to arbitrary parametric models, which can be identified from the second-order spectrum.

Furthermore, algorithmic issues, such as in-order algorithms (e.g., generalizations of the Levinson-Durbin algorithm) need to be developed.

2.5. *Inference for nonparametric tvAR models – An overview*

In the last section, we studied parametric estimates for tvAR(p) models. This is an important option if the length of the time series is short or if we have specific parametric models in mind. In general, however, one would prefer nonparametric models. For nonparametric statistics, a large variety of different estimates are available (local polynomial fits, estimation under shape restrictions, wavelet methods, etc.), and it turns out that it is not too difficult to apply such methods to tvAR(p) models and, moreover, also to other possibly nonlinear models (while the derivation of the corresponding theory may be very challenging). A key role is played by the conditional likelihood at time t which in the tvAR(p) case is

$$\ell_{t,T}(\boldsymbol{\theta}) \coloneqq -\log f_{\boldsymbol{\theta}}\big(X_{t,T}\big|X_{t-1,T},\ldots,X_{1,T}\big) \tag{22}$$

$$= \frac{1}{2}\log\big(2\pi\,\sigma^2\big) + \frac{1}{2\,\sigma^2}\left(X_{t,T} + \sum_{j=1}^{p}\alpha_j\;X_{t-j,T}\right)^2 \tag{23}$$

where $\boldsymbol{\theta} = \big(\alpha_1, \ldots, \alpha_p, \sigma^2\big)'$ and its approximation $\ell^*_{t,T}(\boldsymbol{\theta})$ defined in (96). As a simple example, consider the estimation of the curve $\alpha_1(\cdot)$ of a tvAR(1) process by a local linear fit given by $\hat{\alpha}_1(\cdot) = \hat{c}_0$, where

$$(\hat{c}_0, \hat{c}_1) = \underset{c_0,c_1}{\operatorname{argmin}} \frac{1}{bT} \sum_{t=1}^{T} K\Big(\frac{u_0 - t/T}{b}\Big) \Big(X_{t,T} + \Big[c_0 + c_1\Big(\frac{t}{T} - u_0\Big)\Big] X_{t-1,T}\Big)^2 \quad (24)$$

or more generally (with vectors $\boldsymbol{c}_0$ and $\boldsymbol{c}_1$) given by $\hat{\boldsymbol{\theta}}(u_0) = \hat{\boldsymbol{c}}_0$ with

$$(\hat{\boldsymbol{c}}_0, \hat{\boldsymbol{c}}_1) = \underset{\boldsymbol{c}_0,\boldsymbol{c}_1}{\operatorname{argmin}} \frac{1}{bT} \sum_{t=1}^{T} K\Big(\frac{u_0 - t/T}{b}\Big) \ell_{t,T} \left(\boldsymbol{c}_0 + \boldsymbol{c}_1 \left(\frac{t}{T} - u_0\right)\right). \quad (25)$$

Besides this local linear estimate, many other estimates can be constructed based on the conditional likelihood $\ell_{t,T}(\boldsymbol{\theta})$ from above:

1. A *kernel estimate* defined by

$$\hat{\boldsymbol{\theta}}(u_0) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{bT} \sum_{t=1}^{T} K\Big(\frac{u_0 - t/T}{b}\Big) \ell_{t,T}(\boldsymbol{\theta}). \quad (26)$$

   This estimate is studied in Section 3. We are convinced that it is equivalent to the local Yule-Walker estimate from (7) with $K(x) = h(x)^2$, $b = N/T$ and that all results from (3) are exactly the same for this estimate.
2. A *local polynomial fit* defined by $\hat{\boldsymbol{\theta}}(u_0) = \hat{\boldsymbol{c}}_0$ with

$$(\hat{\boldsymbol{c}}_0, \ldots, \hat{\boldsymbol{c}}_d)' = \underset{\boldsymbol{c}_0,\ldots,\boldsymbol{c}_d}{\operatorname{argmin}} \frac{1}{bT} \sum_{t=1}^{T} K\Big(\frac{u_0 - t/T}{b}\Big) \ell_{t,T} \left(\sum_{j=0}^{d} \boldsymbol{c}_j \left(\frac{t}{T} - u_0\right)^j\right). \quad (27)$$

   Local polynomial fits for tvAR(p) models have been investigated by Kim (2001) and Jentsch (2006).
3. An *orthogonal series estimate* (e.g., a *wavelet estimate*) defined by

$$\bar{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} \ell_{t,T} \left(\sum_{j=1}^{J(T)} \boldsymbol{\beta}_j \psi_j \left(\frac{t}{T}\right)\right) \quad (28)$$

   together with some shrinkage of $\bar{\boldsymbol{\beta}}$ to obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}(u_0) = \sum_{j=1}^{J(T)} \hat{\boldsymbol{\beta}}_j \psi_j(u_0)$. Usually, $J(T) \to \infty$ as $T \to \infty$. Such an estimate has been investigated for a truncated wavelet expansion for tvAR(p) models by Dahlhaus et al. (1999).
4. A *nonparametric maximum likelihood estimate* defined by

$$\hat{\boldsymbol{\theta}}(\cdot) = \underset{\boldsymbol{\theta}(\cdot)\in\Theta}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} \ell_{t,T} \left(\boldsymbol{\theta}\left(\frac{t}{T}\right)\right) \quad (29)$$

where Θ is an adequate function space, for example, a space of curves under shape restrictions such as monotonicity constraints. In the study by Dahlhaus and Polonik (2006), the estimation of a monotonic variance function in a tvAR model is studied, including explicit algorithms involving isotonic regression.

5. A *parametric fit* for the curves $\boldsymbol{\theta}(\cdot) = \boldsymbol{\theta}_\eta(\cdot)$ with $\eta \in \mathbf{R}^q$ defined by

$$\hat{\eta} = \operatorname*{argmin}_\eta \frac{1}{T} \sum_{t=1}^{T} \ell_{t,T}\left(\boldsymbol{\theta}_\eta\left(\frac{t}{T}\right)\right) \tag{30}$$

The resulting estimate has not been investigated yet. It is presumably very close to the exact MLE studied in Theorem 8.

REMARK 3. (i) In the tvAR(p) case, the situation simplifies a lot if $\sigma^2(\cdot) \equiv c$. In that case, the estimates for $\boldsymbol{\alpha}(\cdot)$ and σ^2 "split" and $\ell_{t,T}(\boldsymbol{\theta})$ can in all cases be replaced by $\left(X_{t,T} + \sum_{j=1}^{p} \alpha_j\ X_{t-j,T}\right)^2$ leading to least squares type estimates. (ii) All estimates from above can be transferred to other models by using the conditional likelihood (22) for the specific model. The kernel estimate will be investigated in Section 3. (iii) As mentioned above, an alternative choice is to replace $\ell_{t,T}(\boldsymbol{\theta})$ by the local generalized Whittle likelihood $\ell^*_{t,T}(\boldsymbol{\theta})$ from (96). With that likelihood, several estimates from above have been investigated – see the detailed discussion at the end of Section 5. In that case, the d-dimensional parameter curve $\boldsymbol{\theta}(\cdot) = \left(\theta_1(\cdot), \ldots, \theta_d(\cdot)\right)'$ must be uniquely identifiable from the time varying spectrum $f(u, \lambda) = f_{\boldsymbol{\theta}(u)}(\lambda)$. □

2.6. *Shape and transition curves*

There exist several alternative models for tvAR processes – in particular models where specific characteristics of the time series are modeled by a curve. Below, we give four examples where we restrict ourselves to tvAR(2) models. Suppose, we have a stationary AR(2) model with complex roots $\frac{1}{r}\exp(i\phi)$ and $\frac{1}{r}\exp(-i\phi)$, that is, with parameters $a_1 = -2r\cos(\phi)$, $a_2 = r^2$, and variance σ^2. The corresponding process shows a quasi-periodic behavior with period of length $\frac{2\pi}{\phi}$, that is with frequency ϕ. The more r gets closer to 1, the more the shape of the process gets closer to a sine wave. The amplitude is proportional to σ (if σ (say in (2)) is replaced by $c \cdot \sigma$, then X_t is replaced by $c \cdot X_t$).

In the specific tvAR(2) case, we can now consider the following shape and transition models for quasi-periodic processes:

(i) Model with a time varying amplitude curve:

$$a_1(\cdot),\ a_2(\cdot) \text{ constant}; \quad \sigma(\cdot) \text{ time varying.}$$

Chandler and Polonik (2006) use this model with a unimodal $\sigma(\cdot)$ and a nonparametric maximum likelihood estimate for the discrimination of earthquakes and explosions. The properties of the estimator have been investigated in Dahlhaus and Polonik (2006).

(ii) Model with a time varying frequency curve:

$a_1(\cdot) = -2r\cos\big(\phi(\cdot)\big)$, $a_2(\cdot) = r^2$ with r constant and $\phi(\cdot)$ time varying, $\sigma(\cdot)$ constant.

The model in Fig. 1 is of this form with $r = 0.9$ and $\phi(u) = 1.5 - \cos 4\pi u$.

(iii) Model with a time varying period distinctiveness:

$a_1(\cdot) = -2r(\cdot)\cos(\phi)$, $a_2(\cdot) = r(\cdot)^2$ with $r(\cdot)$ time varying and ϕ constant, $\sigma(\cdot)$ constant.

(iv) Transition models: Amado and Teräsvirta (2011) have recently used the logistic transition function to model parameter transitions in GARCH models. The simplest transition function is

$$G\Big(\frac{t}{T};\gamma,c\Big) := \Big[1+\exp\Big\{-\gamma\Big(\frac{t}{n}-c\Big)\Big\}\Big]^{-1}.$$

Since $G(0;\gamma,c) \approx 0$ and $G(1;\gamma,c) \approx 1$, the model

$$\begin{aligned} a_1(u) &= a_1^{\text{start}} + G(u;\gamma,c)\big(a_1^{\text{end}} - a_1^{\text{start}}\big),\ a_2(u) \\ &= a_2^{\text{start}} + G(u;\gamma,c)\big(a_2^{\text{end}} - a_2^{\text{start}}\big) \end{aligned}$$

is a parametric model for a smooth transition from the AR model with parameters $(a_1^{\text{start}}, a_2^{\text{start}})$ at $u = 0$ to the model with parameters $(a_1^{\text{end}}, a_2^{\text{end}})$ at $u = 1$. Here, c and γ are the location and the "smoothness" of transition, respectively. More general transition models (in particular with more states) may be found in the study by Amado and Teräsvirta (2011). $G(\cdot;\gamma,c)$ may also be replaced by a (nonparametric) function $G(\cdot)$ with $G(0) = 0$ and $G(1) = 0$.

It is obvious that all methods from subsection 5 can be applied in cases (i)–(iv) to estimate the constant parameters and the shape and transition curves. We mention that the theoretical results for local Whittle estimates of Dahlhaus and Giraitis (1998) apply to these models (cf. Example 3); the uniform convergence result for the local generalized Whittle estimate in Theorem 13; the asymptotic results of Dahlhaus and Neumann (2001), where the parameter curves are estimated by a nonlinear wavelet method; the results of Dahlhaus and Polonik (2006) on nonparametric maximum likelihood estimates under shape constraints; and the results for parametric models in Theorem 8 on the MLE and the generalized Whittle estimator and by Dahlhaus (1997) on the block Whittle estimator.

3. Local likelihoods, derivative processes, and nonlinear models with time varying parameters

In this section, we present a more general framework for time series with time varying finite-dimensional parameters $\boldsymbol{\theta}(\cdot)$ and show how nonparametric inference can be

done and theoretically handled. Typically, such models result from the generalization of classical parametric models to the time varying case. If we restrict ourselves to linear processes or even more to Gaussian processes, then a much more general theory is possible which is developed in the subsequent sections. Large parts of the present section are based on the ideas presented by Dahlhaus and Subba Rao (2006), where time varying ARCH models have been investigated.

The key idea is to use at each time point $u_0 \in (0,1)$, the stationary approximation $\tilde{X}_t(u_0)$ to the original process $X_{t,T}$, and to calculate the bias resulting from the use of this approximation. This will end in Taylor-type expansions of $X_{t,T}$ in terms of so-called derivative processes. These expansions play a major role in the theoretical derivations.

Suppose for example that we estimate the multivariate parameter curve $\boldsymbol{\theta}(\cdot)$ by minimizing the (negative) local conditional log-likelihood, that is

$$\hat{\boldsymbol{\theta}}_T^C(u_0) := \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_T^C(u_0,\boldsymbol{\theta})$$

with

$$\mathcal{L}_T^C(u_0,\boldsymbol{\theta}) := \frac{1}{T}\sum_{t=1}^{T} \frac{1}{b} K\left(\frac{u_0 - t/T}{b}\right) \ell_{t,T}(\boldsymbol{\theta}) \tag{31}$$

and

$$\ell_{t,T}(\boldsymbol{\theta}) := -\log f_{\boldsymbol{\theta}}\left(X_{t,T}\middle| X_{t-1,T},\ldots,X_{1,T}\right)$$

where K is symmetric, has compact support $[-\frac{1}{2},\frac{1}{2}]$, and fulfills $\int_{-1/2}^{1/2} K(x)\,dx = 1$. We assume that $b = b_T \to 0$ and $b\,T \to \infty$ as $T \to \infty$. Two examples for this likelihood are given below.

We approximate $\mathcal{L}_T^C(u_0,\boldsymbol{\theta})$ with $\tilde{\mathcal{L}}_T^C(u_0,\boldsymbol{\theta})$, which is the same function but with $\ell_{t,T}(\boldsymbol{\theta})$ replaced by

$$\tilde{\ell}_t(u_0,\boldsymbol{\theta}) := -\log f_{\boldsymbol{\theta}}\big(\tilde{X}_t(u_0)\big|\tilde{X}_{t-1}(u_0),\ldots,\tilde{X}_1(u_0)\big),$$

which means that $X_{t,T}$ is replaced by its stationary approximation $\tilde{X}_t(u_0)$. Usually, this is the local conditional likelihood for the process $\tilde{X}_t(u_0)$.

Example 1. (i) Consider the tvAR(p) process defined in (2) together with its stationary approximation at time u_0 given by (3). Under suitable regularity conditions, it can be shown that $X_{t,T} = \tilde{X}_t(u_0) + O_p\big(\big|\frac{t}{T} - u_0\big| + \frac{1}{T}\big)$ (cf. (51)). In case where the ε_t are Gaussian, the conditional likelihood at time t is given by

$$\ell_{t,T}(\boldsymbol{\theta}) = \frac{1}{2}\log\big(2\pi\,\sigma^2\big) + \frac{1}{2\,\sigma^2}\left(X_{t,T} + \sum_{j=1}^{p}\alpha_j\, X_{t-j,T}\right)^2 \tag{32}$$

where $\boldsymbol{\theta} = (\alpha_1,\ldots,\alpha_p,\sigma^2)'$. It is easy to show that the resulting estimate is the same as in (7) but with $\hat{r}_T(u_0) := \big(\tilde{\tilde{c}}_T(u_0,0,1),\ldots,\tilde{\tilde{c}}_T(u_0,0,p)\big)'$ and $\hat{R}_T(u_0) := \{\tilde{\tilde{c}}_T(u_0,i,j)\}_{i,j=1,\ldots,p}$ with the local covariance estimator $\tilde{\tilde{c}}_T(u,i,j)$ as defined in (10).

(ii) A tvARCH(p) model where $\{X_{t,T}\}$ is assumed to satisfy the representation

$$X_{t,T} = \sigma_{t,T} Z_t$$

$$\text{where} \quad \sigma_{t,T}^2 = \alpha_0\left(\frac{t}{T}\right) + \sum_{j=1}^{p} \alpha_j\left(\frac{t}{T}\right) X_{t-j,N}^2 \quad \text{for } t = 1, \ldots, N \tag{33}$$

with Z_t being independent, identically distributed random variables with $\mathbf{E}Z_t = 0$, $\mathbf{E}Z_t^2 = 1$.

The corresponding stationary approximation $\tilde{X}_t(u_0)$ at time u_0 is given by

$$\tilde{X}_t(u_0) = \sigma_t(u_0)\, Z_t$$

$$\text{where} \quad \sigma_t(u_0)^2 = \alpha_0(u_0) + \sum_{j=1}^{p} \alpha_j(u_0)\, \tilde{X}_{t-j}(u_0)^2 \quad \text{for } t \in \mathbf{Z}. \tag{34}$$

It is shown by Dahlhaus and Subba Rao (2006) that $\{X_{t,T}^2\}$ as defined above has an almost surely well-defined unique solution in the set of all causal solutions and $X_{t,T}^2 = \tilde{X}_t(u_0)^2 + O_p(|\frac{t}{T} - u_0| + \frac{1}{N})$. In case where the Z_t are Gaussian, the conditional likelihood is given by

$$\ell_{t,T}(\boldsymbol{\theta}) = \frac{1}{2}\log w_{t,T}(\boldsymbol{\theta}) + \frac{X_{t,T}^2}{2\, w_{t,T}(\boldsymbol{\theta})} \text{ with } w_{t,T}(\boldsymbol{\theta}) = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j,T}^2 \tag{35}$$

where $\boldsymbol{\theta} = (\alpha_0, \ldots, \alpha_p)'$. Dahlhaus and Subba Rao (2006) prove consistency of the resulting estimate also in case where the true process is not Gaussian. As an alternative, Fryzlewicz et al. (2008) propose a kernel normalized-least-squares estimator that has a closed form and thus has some advantages over the above kernel estimate for small samples.

(iii) Another example is a tvGARCH(p,q) process – see Example 6.

We now discuss the derivation of the asymptotic bias, mean-squared error, consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}_T(u_0)$ for an "arbitrary" local minimum-distance function $\mathcal{L}_T(u_0, \boldsymbol{\theta})$ (keeping in mind the above local conditional likelihood). The results are obtained by approximating $\mathcal{L}_T(u_0, \boldsymbol{\theta})$ with $\tilde{\mathcal{L}}_T(u_0, \boldsymbol{\theta})$, which is the same function but with $X_{t,T}$ replaced by its stationary approximation $\tilde{X}_t(u_0)$. Typically, both $\mathcal{L}_T(u_0, \boldsymbol{\theta})$ and $\tilde{\mathcal{L}}_T(u_0, \boldsymbol{\theta})$ will converge to the same limit-function, which we denote by $\mathcal{L}(u_0, \boldsymbol{\theta})$. Let

$$\boldsymbol{\theta}_0(u_0) \coloneqq \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(u_0, \boldsymbol{\theta}).$$

If the model is correctly specified, then, typically, $\boldsymbol{\theta}_0(u_0)$ is the true curve. Furthermore, let

$$\mathcal{B}_T(u_0, \boldsymbol{\theta}) \coloneqq \mathcal{L}_T(u_0, \boldsymbol{\theta}) - \tilde{\mathcal{L}}_T(u_0, \boldsymbol{\theta}).$$

The following two results describe how the asymptotic properties of $\hat{\boldsymbol{\theta}}_T(u_0)$ can be derived. They should be regarded as a general road map, and the challenge is to prove the conditions in a specific situation which may be quite difficult.

THEOREM 2. *(i) Suppose that Θ is compact with $\boldsymbol{\theta}_0(u_0) \in Int(\Theta)$, the function $\mathcal{L}(u_0, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and the minimum $\boldsymbol{\theta}_0(u_0)$ is unique. If*

$$\sup_{\boldsymbol{\theta}\in\Theta} \left|\tilde{\mathcal{L}}_T(u_0, \boldsymbol{\theta}) - \mathcal{L}(u_0, \boldsymbol{\theta})\right| \xrightarrow{P} 0, \tag{36}$$

and

$$\sup_{\boldsymbol{\theta}\in\Theta} \left|\mathcal{B}_T(u_0, \boldsymbol{\theta})\right| \xrightarrow{P} 0 \tag{37}$$

then

$$\hat{\boldsymbol{\theta}}_T(u_0) \xrightarrow{P} \boldsymbol{\theta}_0(u_0). \tag{38}$$

(ii) Suppose in addition that $\mathcal{L}(u, \boldsymbol{\theta})$ and $\boldsymbol{\theta}_0(u)$ are uniformly continuous in u and $\boldsymbol{\theta}$, and the convergence in (36) and (37) is uniformly in $u_0 \in [0, 1]$. Then

$$\sup_{u_0\in[0,1]} \left|\hat{\boldsymbol{\theta}}_T(u_0) - \boldsymbol{\theta}_0(u_0)\right| \xrightarrow{P} 0. \tag{39}$$

PROOF. The proof of (i) is standard – cf. the proof of Theorem 2 in the study by Dahlhaus and Subba Rao (2006). The proof of (ii) is a straightforward generalization. □

Note that in (i), all conditions apart from (37) are conditions on the stationary process $\tilde{X}_t(u_0)$ with (fixed) parameter $\boldsymbol{\theta}(u_0)$ and the stationary likelihood/minimum-distance function $\tilde{\mathcal{L}}_T(u_0, \boldsymbol{\theta})$. These properties are usually known from existing results on stationary processes. It only remains to verify the condition (37), which can be done by using the expansion (51) in terms of derivative processes (see the discussion below). (ii) contains a little pitfall: Usually, the estimate $\hat{\boldsymbol{\theta}}_T(u_0)$ is defined for $u_0 = 0$ or $u_0 = 1$ in a different way due to edge effects. This means that also $\tilde{\mathcal{L}}_T(u_0, \boldsymbol{\theta})$ looks different, that is, one would usually prefer a uniform convergence result for $u_0 \in (0, 1)$, which is more difficult to prove.

Even more interesting and challenging is a uniform convergence result with a rate of convergence. For time varying AR(p) processes, this is stated for a different likelihood in Theorem 13. We mention that such a result usually requires an exponential bound and maximal inequalities which need to be tailored to the specific model at hand.

We now state the corresponding result on asymptotic normality in case of second-order smoothness. ∇ denotes the derivatives with respect to the θ_i, i.e., $\nabla := \left(\partial/\partial\theta_i\right)_{i=1,\ldots,d}$.

THEOREM 3. *Let* $\boldsymbol{\theta}_0 := \boldsymbol{\theta}_0(u_0)$. *Suppose that* $\mathcal{L}_T(u_0, \boldsymbol{\theta})$, $\tilde{\mathcal{L}}_T(u_0, \boldsymbol{\theta})$, *and* $\mathcal{L}(u_0, \boldsymbol{\theta})$ *are twice continuously differentiable in* $\boldsymbol{\theta}$ *with nonsingular matrix* $\Gamma(u_0) := \nabla^2 \mathcal{L}(u_0, \boldsymbol{\theta}_0)$. *Let further*

$$\sqrt{b\,T}\ \nabla \tilde{\mathcal{L}}_T(u_0, \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, V(u_0)\big)$$

with some sequence $b = b_T$, *where* $b \to 0$ *and* $b\,T \to \infty$ *(the definition of* b *is part of the definition of the likelihood – it is usually some bandwidth) and*

$$\sup_{\boldsymbol{\theta} \in \Theta} \left| \nabla^2 \tilde{\mathcal{L}}_T(u_0, \boldsymbol{\theta}) - \nabla^2 \mathcal{L}(u_0, \boldsymbol{\theta}) \right| \xrightarrow{P} 0.$$

If in addition

$$\sqrt{bT}\Big(\Gamma(u_0)^{-1}\, \nabla \mathcal{B}_T(u_0, \boldsymbol{\theta}_0) - \frac{b^2}{2}\, \boldsymbol{\mu}^0(u_0)\Big) = o_p(1) \tag{40}$$

with some $\boldsymbol{\mu}^0(\cdot)$ *(to be specified below – cf. (47)) and*

$$\sup_{\boldsymbol{\theta} \in \Theta} \left| \nabla^2 \mathcal{B}_T(u_0, \boldsymbol{\theta}) \right| \xrightarrow{P} 0 \tag{41}$$

then

$$\sqrt{b\,T}\ \Big(\hat{\boldsymbol{\theta}}_T(u_0) - \boldsymbol{\theta}_0(u_0) + \frac{b^2}{2} \boldsymbol{\mu}^0(u_0)\Big) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \Gamma(u_0)^{-1}\, V(u_0)\, \Gamma(u_0)^{-1}\Big). \tag{42}$$

PROOF. The usual Taylor expansion of $\nabla \mathcal{L}_T(u_0, \boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$ yields

$$\sqrt{b\,T}\ \big(\hat{\boldsymbol{\theta}}_T(u_0) - \boldsymbol{\theta}_0 + \Gamma(u_0)^{-1} \nabla \mathcal{B}_T(u_0, \boldsymbol{\theta}_0)\big) = -\sqrt{b\,T}\ \Gamma(u_0)^{-1}\, \nabla \tilde{\mathcal{L}}_T(u_0, \boldsymbol{\theta}_0) + o_p(1). \tag{43}$$

The result then follows immediately. □

REMARK 4. (i) Again the first two conditions are conditions on the stationary process $\tilde{X}_t(u_0)$ with (fixed) parameter $\boldsymbol{\theta}(u_0)$ and the stationary likelihood/minimum-distance function $\tilde{\mathcal{L}}_T(u_0, \boldsymbol{\theta})$, which are usually known from existing results on stationary processes.

(ii) Of course, an analogous result also holds under different smoothness conditions and with other rates than b^2 in (40) and (42).

(iii) Under additional regularity conditions, one can usually prove that the same expansion as in (43) also holds for the moments, leading to

$$\mathbf{E}\, \hat{\boldsymbol{\theta}}_T(u_0) = \boldsymbol{\theta}_0(u_0) - \frac{b^2}{2}\, \boldsymbol{\mu}^0(u_0) + o(b^2) \tag{44}$$

and

$$\operatorname{var}\big(\hat{\boldsymbol{\theta}}_T(u_0)\big) = \frac{1}{bT}\,\Gamma(u_0)^{-1}\,V(u_0)\,\Gamma(u_0)^{-1} + o\Big(\frac{1}{bT}\Big) \tag{45}$$

(note that (43) is a stochastic expansion which does not automatically imply these moment relations). The proof of these properties is usually not easy. □

Example 2 (Kernel-type local likelihoods). We now return to the local conditional likelihood (31) as a special case and provide some heuristics on how to calculate the above terms (in particular, the bias $\boldsymbol{\mu}^0(u_0)$). We stress that in the concrete situation, where a specific model is given the exact proof usually goes along the same lines but the details may be quite challenging.

Suppose that the local likelihood of the stationary process $\tilde{X}_t(u_0)$ converges in probability to

$$\mathcal{L}(u_0,\boldsymbol{\theta}) := \lim_{T\to\infty} \tilde{\mathcal{L}}_T(u_0,\boldsymbol{\theta}) = \lim_{t\to\infty} \mathbf{E}\,\tilde{\ell}_t(u_0,\boldsymbol{\theta}).$$

Usually, we have $X_{t,T} = \tilde{X}_t(t/T) + O_p(T^{-1})$ and

$$\mathbf{E}\,\nabla\ell_{t,T}\,(\boldsymbol{\theta}) = \mathbf{E}\,\nabla\tilde{\ell}_t\left(\frac{t}{T},\boldsymbol{\theta}\right) + o\big((bT)^{-1/2}\big) = \nabla\mathcal{L}\left(\frac{t}{T},\boldsymbol{\theta}\right) + o\big((bT)^{-1/2}\big)$$

uniformly in t. A Taylor expansion then leads in the case $b^3 = o((bT)^{-1/2})$ with the symmetry of the kernel K to

$$\begin{aligned}
\mathbf{E}\,\nabla\mathcal{L}_T(u_0,\boldsymbol{\theta}) &= \frac{1}{bT}\sum_{t=1}^{T} K\left(\frac{u_0 - t/T}{b}\right)\nabla\mathcal{L}\left(\frac{t}{T},\boldsymbol{\theta}\right) + o\big((bT)^{-1/2}\big)\\
&= \nabla\mathcal{L}(u_0,\boldsymbol{\theta}) + \Big[\frac{\partial}{\partial u}\nabla\mathcal{L}(u_0,\boldsymbol{\theta})\Big]\frac{1}{bT}\sum_{t=1}^{T} K\Big(\frac{u_0 - t/T}{b}\Big)\Big(\frac{t}{T} - u_0\Big)\\
&\quad + \frac{1}{2}\Big[\frac{\partial^2}{\partial u^2}\nabla\mathcal{L}(u_0,\boldsymbol{\theta})\Big]\frac{1}{bT}\sum_{t=1}^{T} K\Big(\frac{u_0 - t/T}{b}\Big)\Big(\frac{t}{T} - u_0\Big)^2\\
&\quad + o\big((bT)^{-1/2}\big)\\
&= \nabla\mathcal{L}(u_0,\boldsymbol{\theta}) + \frac{1}{2}\,b^2\,d_K\,\frac{\partial^2}{\partial u^2}\nabla\mathcal{L}(u_0,\boldsymbol{\theta}) + o\big((bT)^{-1/2}\big)
\end{aligned} \tag{46}$$

with $d_K := \int x^2 K(x)\,dx$. Since $\mathbf{E}\,\nabla\tilde{\mathcal{L}}_T(u_0,\boldsymbol{\theta}) = \nabla\mathcal{L}(u_0,\boldsymbol{\theta}) + o\big((bT)^{-1/2}\big)$, this leads with (40) to the bias term

$$\boldsymbol{\mu}^0(u_0) = d_K\,\Gamma(u_0)^{-1}\,\frac{\partial^2}{\partial u^2}\nabla\mathcal{L}\big(u,\boldsymbol{\theta}_0(u_0)\big)\big]_{u=u_0} =: d_k\,\boldsymbol{\mu}(u_0) \tag{47}$$

Let $\boldsymbol{\theta}_0 \coloneqq \boldsymbol{\theta}_0(u_0)$. If the model is correctly specified, it usually can be shown that $\nabla\tilde{\ell}_t(u_0, \boldsymbol{\theta}_0)$ is a martingale difference sequence, and the condition of the Lindeberg martingale central limit theorem is fulfilled leading to

$$\sqrt{b\,T}\;\nabla\tilde{\mathcal{L}}_T(u_0, \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, v_K\,\mathbf{E}\big(\nabla\tilde{\ell}_t(u_0, \boldsymbol{\theta}_0)\big)\big(\nabla\tilde{\ell}_t(u_0, \boldsymbol{\theta}_0)\big)'\Big)$$

with $v_K = \int K(x)^2 dx$. Furthermore, if the model is correctly specified, we usually have

$$\mathbf{E}\big(\nabla\tilde{\ell}_t(u_0, \boldsymbol{\theta}_0)\big)\big(\nabla\tilde{\ell}_t(u_0, \boldsymbol{\theta}_0)\big)' = \nabla^2\mathcal{L}(u_0, \boldsymbol{\theta}_0) = \Gamma(u_0)$$

that is

$$\sqrt{b\,T}\;\Big(\hat{\boldsymbol{\theta}}_T(u_0) - \boldsymbol{\theta}_0(u_0) + \frac{b^2}{2}\,d_K\,\Gamma(u_0)^{-1}\,\frac{\partial^2}{\partial u^2}\nabla\mathcal{L}(u_0, \boldsymbol{\theta}_0)\Big) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, v_K\,\Gamma(u_0)^{-1}\Big). \tag{48}$$

If we are able to prove, in addition, the formulas (44) and (45) on the asymptotic bias and variance, we obtain the same formula for the asymptotic mean-squared error as in (11) with $\tau(u_0) = \mathrm{tr}\{\Gamma(u_0)^{-1}\}$ and $\mu(u_0)^2$ replaced by $\|\boldsymbol{\mu}(u_0)\|^2$, where $\boldsymbol{\mu}(u_0) = \Gamma(u_0)^{-1}\,\frac{\partial^2}{\partial u^2}\nabla\mathcal{L}(u_0, \boldsymbol{\theta}_0)$. As in Remark 1, this leads to the optimal segment length and the optimal mean-squared error. The implications for nonrescaled processes are the same as in Remark 2.

We now present three examples where the above results have been proved explicitly.

Example 3 (Local Whittle estimates). The first example are local Whittle estimates on segments $\hat{\boldsymbol{\theta}}_T^W(u_0)$ obtained by minimizing $\mathcal{L}_T^W(u_0, \boldsymbol{\theta})$ (cf. (18)). In case of a tvAR(p)process, $\hat{\boldsymbol{\theta}}_T^W(u_0)$ is exactly the local Yule-Walker estimate defined in (7) with the covariance-estimates given in (8). $\mathcal{L}_T^W(u, \boldsymbol{\theta})$ is not exactly a local conditional likelihood as defined in (31), but approximately (in the same sense as $\hat{c}_T(u_0, k)$ from (8) is an approximation to the kernel covariance estimate). For that reason, the above heuristics also applies to this estimate and can be made rigorous.

In Dahlhaus and Giraitis (1998), Theorems 3.1 and 3.2 bias and asymptotic normality of $\hat{\boldsymbol{\theta}}_T^W(u_0)$ have been derived rigorously including a derivation of the variance and the mean-squared error as given in (44) and (45) (i.e., not only the stochastic expansion in (43)). We mention that, therefore, also the results on the optimal kernel and bandwidth in (12) and (13) apply to this situation.

In the present situation, we have (cf. Dahlhaus and Giraitis, 1998, (3.7))

$$\mathcal{L}(u, \boldsymbol{\theta}) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\Big\{\log 4\pi^2 f_{\boldsymbol{\theta}}(\lambda) + \frac{f(u,\lambda)}{f_{\boldsymbol{\theta}}(\lambda)}\Big\}\,d\lambda.$$

Therefore,

$$\frac{\partial^2}{\partial u^2}\nabla\mathcal{L}(u_0, \boldsymbol{\theta}) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\nabla f_{\boldsymbol{\theta}}(\lambda)^{-1}\,\frac{\partial^2}{\partial u^2} f(u_0, \lambda)\,d\lambda$$

and in the correctly specified case where $f(u,\lambda) = f_{\boldsymbol{\theta}_0(u)}(\lambda)$

$$\Gamma(u_0) = \nabla^2 \mathcal{L}(u_0, \boldsymbol{\theta}_0) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \big(\nabla \log f_{\boldsymbol{\theta}_0}\big)\big(\nabla \log f_{\boldsymbol{\theta}_0}\big)' \, d\lambda$$

leading to the asymptotic bias $\boldsymbol{\mu}(u_0)$ in (47) and the asymptotic variance in the central limit theorem (48). A uniform convergence result for $\hat{\boldsymbol{\theta}}_T^W(u_0)$ is stated in Theorem 13.

Example 4 (tvAR(p) processes). In the special case of a Gaussian tvAR(p) process, the exact results for the local Yule-Walker estimates (7) follow as a special case from the above results on local Whittle estimates (see also Section 2 in Dahlhaus and Giraitis (1998), where tvAR(p) processes are discussed separately). In that case, we have with $R(u)$ and $r(u)$ as in (6) that $\Gamma(u) = 1/\sigma^2(u)\, R(u)$. Furthermore,

$$\nabla \mathcal{L}\big(u, \boldsymbol{\theta}\big) = \frac{1}{\sigma^2}\big[R(u)\,\boldsymbol{\alpha} + r(u)\big]$$

which implies

$$\boldsymbol{\mu}(u_0) = R(u_0)^{-1}\left[\Big(\frac{\partial^2}{\partial u^2} R(u)\Big)\boldsymbol{\alpha}(u_0) + \Big(\frac{\partial^2}{\partial u^2} r(u)\Big)\right]_{u=u_0}.$$

We conjecture that exactly the same asymptotic results hold for the conditional likelihood estimate obtained by minimizing

$$\mathcal{L}_T^C(u_0, \boldsymbol{\theta}) := \frac{1}{T}\sum_{t=1}^{T} \frac{1}{b} K\Big(\frac{u_0 - t/T}{b}\Big)\left[\frac{1}{2}\log\big(2\pi\,\sigma^2\big) + \frac{1}{2\,\sigma^2}\Big(X_{t,T} + \sum_{j=1}^{p} \alpha_j\, X_{t-j,T}\Big)^2\right].$$

We now introduce derivative processes. The key idea in the proofs of Dahlhaus and Giraitis (1998) is to use at time $u_0 \in (0,1)$, the stationary approximation $\tilde{X}_t(u_0)$ (there denoted by Y_t) to the original process $X_{t,T}$ and to calculate the bias resulting from the use of this approximation. As in Dahlhaus and Subba Rao (2006), we now extend this idea leading to the Taylor-type expansion (51), which is an expansion of the original process in terms of (usually ergodic) stationary processes called derivative processes. This expansion is a powerful tool, since all techniques for stationary processes including the ergodic theorem may be applied for the local investigation of the nonstationary process $X_{t,T}$. The use of this expansion and of derivative processes in general leads to a general structure of the proofs and simplifies the derivations a lot.

We start with the simple example of a tvAR(1) process, since in this case everything can be calculated directly. Then, $X_{t,T}$ is defined by $X_{t,T} + \alpha_1(t/T)X_{t-1,T} = \varepsilon_t, t \in \mathbf{Z}$ and the stationary approximation $\tilde{X}_t(u_0)$ at time $u_0 = t_0/n$ by $\tilde{X}_t(u_0) +$

$\alpha_1(u_0)\tilde{X}_t(u_0) = \varepsilon_t, t \in \mathbf{Z}$. Repeated plug-in yields under suitable regularity conditions (for a rigorous argument, see the proof of Theorem 2.3 in Dahlhaus (1996a)),

$$X_{t,T} = \sum_{j=0}^{\infty}(-1)^j \Big[\prod_{k=0}^{j-1}\alpha_1\Big(\frac{t-k}{T}\Big)\Big]\varepsilon_{t-j} = \sum_{j=0}^{\infty}(-1)^j\,\alpha_1\Big(\frac{t}{T}\Big)^j\,\varepsilon_{t-j} + O_p\Big(\frac{1}{T}\Big) \quad (49)$$

$$= \tilde{X}_t\Big(\frac{t}{T}\Big) + O_p\Big(\frac{1}{T}\Big) = \tilde{X}_t(u_0) + \Big(\frac{t}{T} - u_0\Big)\frac{\partial \tilde{X}_t(u)}{\partial u}\rfloor_{u=u_0} + O_p\Big(\frac{1}{T}\Big). \quad (50)$$

We have in the present situation

$$\frac{\partial \tilde{X}_t(u)}{\partial u} = \sum_{j=0}^{\infty}(-1)^j\,\frac{\partial \alpha_1(u)^j}{\partial u}\,\varepsilon_{t-j} = \sum_{j=0}^{\infty}(-1)^j\big[j\,\alpha_1(u)^{j-1}\alpha_1(u)'\big]\,\varepsilon_{t-j}$$

that is, $\partial\tilde{X}_t(u)/\partial u$ is a stationary ergodic process in t with $\left|\partial\tilde{X}_t(u)/\partial u\right| \le \sum_{j=1}^{\infty} j\rho^{j-1}|\varepsilon_{t-j}|$, where $|\rho| < 1$. In the same way, we have

$$X_{t,T} = \tilde{X}_t(u_0) + \Big(\frac{t}{T} - u_0\Big)\frac{\partial \tilde{X}_t(u)}{\partial u}\rfloor_{u=u_0} + \frac{1}{2}\Big(\frac{t}{T} - u_0\Big)^2\frac{\partial^2 \tilde{X}_t(u)}{\partial u^2}\rfloor_{u=u_0} + O_p\left(\Big(\frac{t}{T} - u_0\Big)^3 + \frac{1}{T}\right) \quad (51)$$

with the second-order derivative process $\partial^2\tilde{X}_t(u)/\partial u^2\rfloor_{u=u_0}$, which is defined analogously. It is not difficult to prove existence and uniqueness in a rigorous sense.

For general tvAR(p) processes, the same results holds – however, it is difficult in that case to write the derivative process in explicit form. It is interesting to note that the derivative process fulfills the equation

$$\frac{\partial \tilde{X}_t(u)}{\partial u} + \sum_{j=1}^{p}\Big(\alpha_j(u)\,\frac{\partial \tilde{X}_{t-j}(u)}{\partial u} + \alpha_j'(u)\,\tilde{X}_{t-j}(u)\Big) = \frac{\partial \sigma(u)}{\partial u}\,\varepsilon_t,$$

where $\alpha_j'(u)$ denotes the derivative of $\alpha_j(u)$ with respect to u. This is formally obtained by differentiating both sides of Eq. (3). Furthermore, it can be shown that this equation system uniquely defines the derivative process.

We are convinced that the expansion (51) and equation systems like (52) can be established for several other locally stationary time series models. As mentioned above, the important point is that (51) is an expansion in terms of stationary processes.

In the next example, we show how derivative processes are used for deriving the properties of local likelihood estimates.

Example 5 (tvARCH processes). The definition of the processes $X_{t,T}$ and $\tilde{X}_t(u_0)$ has been given above in (33) and (34) and of the local likelihood in (35) and (31). In the study by Dahlhaus and Subba Rao (2006), Theorems 2 and 3, consistency and asymptotic normality, have been established for the resulting estimate, and in

particular, (48) has been proved. Derivative processes play a major role in the proofs, and we briefly indicate how they are used. First, existence and uniqueness of the derivative processes have been proved including the Taylor-type expansion for the process $X_{t,T}^2$:

$$X_{t,T}^2 = \tilde{X}_t(u_0)^2 + \Big(\frac{t}{T} - u_0\Big) \frac{\partial \tilde{X}_t(u)^2}{\partial u}\rfloor_{u=u_0} + \frac{1}{2}\Big(\frac{t}{T} - u_0\Big)^2 \frac{\partial^2 \tilde{X}_t(u)^2}{\partial u^2}\rfloor_{u=u_0} + O_p\bigg(\Big(\frac{t}{T} - u_0\Big)^3 + \frac{1}{T}\bigg) \tag{52}$$

(in this model, we are working with $X_{t,T}^2$ rather than $X_{t,T}$ since $X_{t,T}^2$ is uniquely determined). Furthermore, $\partial \tilde{X}_t(u)^2/\partial u$ is almost surely the unique solution of the equation

$$\frac{\partial \tilde{X}_t(u)^2}{\partial u} = \bigg(\alpha_0'(u) + \sum_{j=1}^{\infty} \alpha_j'(u)\, \tilde{X}_{t-j}(u)^2 + \sum_{j=1}^{\infty} \alpha_j(u)\, \frac{\partial \tilde{X}_{t-j}(u)^2}{\partial u}\bigg)\, Z_t^2 \tag{53}$$

which can formally be obtained by differentiating (34). By taking the second derivative of this expression, we obtain a similar expression for the second derivative $\partial^2 \tilde{X}_t(u)^2/\partial u^2$ etc.

A key step in the above proofs is the derivation of (40) and of the bias term $\boldsymbol{\mu}^0(\cdot)$ in this situation. We briefly sketch this. We have with $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0(u_0)$

$$\nabla \mathcal{B}_T(u_0, \boldsymbol{\theta}_0) = \frac{1}{bT} \sum_{t=1}^{T} K\Big(\frac{u_0 - t/T}{b}\Big) \big(\nabla \ell_{t,T}(\boldsymbol{\theta}_0) - \nabla \tilde{\ell}_t(u_0, \boldsymbol{\theta}_0)\big).$$

First, $\nabla \ell_{t,T}(\boldsymbol{\theta}_0)$ is replaced by $\nabla \tilde{\ell}_t(t/T, \boldsymbol{\theta}_0)$, where we omit details (this works since $X_{t,T}^2$ is approximately the same as $\tilde{X}_t^2(t/T)$). Then, a Taylor-expansion is applied:

$$\begin{aligned} \nabla \tilde{\ell}_t\Big(\frac{t}{T}, \boldsymbol{\theta}_0\Big) - \nabla \tilde{\ell}_t(u_0, \boldsymbol{\theta}_0) &= \Big(\frac{t}{T} - u_0\Big) \frac{\partial \nabla \tilde{\ell}_t(u, \boldsymbol{\theta}_0)}{\partial u}\rfloor_{u=u_0} \\ &\quad + \frac{1}{2}\Big(\frac{t}{T} - u_0\Big)^2 \frac{\partial^2 \nabla \tilde{\ell}_t(u, \boldsymbol{\theta}_0)}{\partial u^2}\rfloor_{u=u_0} \\ &\quad + \frac{1}{6}\Big(\frac{t}{T} - u_0\Big)^3 \frac{\partial^3 \nabla \tilde{\ell}_t(u, \boldsymbol{\theta}_0)}{\partial u^3}\rfloor_{u=\tilde{U}_t} \end{aligned} \tag{54}$$

with a random variable $\tilde{U}_t \in (0, 1]$. The breakthrough now is that $\partial \nabla \tilde{\ell}_t(u, \boldsymbol{\theta}_0)/\partial u$ can be written explicitly in terms of the derivative process $\partial \tilde{X}_t(u)^2/\partial u$ and of the process $\tilde{X}_t(u)^2$, that is, we obtain with the formula for the total derivative

$$\frac{\partial \nabla \tilde{\ell}_t(u, \boldsymbol{\theta}_0)}{\partial u} = \sum_{j=0}^{p} \bigg(\frac{\partial}{\partial \tilde{X}_{t-j}(u)^2}\bigg[\frac{\nabla w_t(u, \boldsymbol{\theta}_0)}{w_t(u, \boldsymbol{\theta}_0)} - \frac{\tilde{X}_t(u)^2 \nabla w_t(u, \boldsymbol{\theta}_0)}{w_t(u, \boldsymbol{\theta}_0)^2}\bigg] \times \frac{\partial \tilde{X}_{t-j}(u)^2}{\partial u}\bigg),$$

where $w_t(u, \boldsymbol{\theta}) = c_0(\boldsymbol{\theta}_0) + \sum_{j=1}^{\infty} c_j(\boldsymbol{\theta}) \tilde{X}_{t-j}(u)^2$ (the same holds true for the higher-order terms). In particular, $\partial \nabla \tilde{\ell}_t(u, \boldsymbol{\theta}_0)/\partial u$ is a stationary process with constant

mean. Due to the symmetry of the kernel, we, therefore, obtain after some lengthy but straightforward calculations

$$\sqrt{bT}\Big(\Gamma(u_0)^{-1}\,\nabla\mathcal{B}_T(u_0,\boldsymbol{\theta}_0) - \frac{b^2}{2}\,d_K\,\Gamma(u_0)^{-1}\,\frac{\partial^2}{\partial u^2}\nabla\mathcal{L}\big(u,\boldsymbol{\theta}_0\big)\rfloor_{u=u_0}\Big) = o_p(1). \quad (55)$$

A very simple example is the tvARCH(0) process

$$X_{t,T} = \sigma_{t,T}\,Z_t, \quad \sigma^2_{t,T} = \alpha_0\left(\frac{t}{T}\right).$$

In this case, $\frac{\partial \bar{X}_t(u)^2}{\partial u} = \alpha_0'(u)\,Z_t^2$ and we have

$$\frac{\partial^2\nabla\mathcal{L}(u,\boldsymbol{\alpha}_{u_0})}{\partial u^2}\rfloor_{u=u_0} = -\frac{1}{2}\frac{\alpha_0''(u_0)}{\alpha_0(u_0)^2} \quad \text{and} \quad \Sigma(u_0) = \frac{1}{2\,\alpha_0(u_0)^2}$$

that is, $\mu(u_0) = -\alpha_0''(u_0)$. This is another example which illustrates how the bias is linked to the nonstationarity of the process - if the process were stationary the derivatives of $\alpha_0(\cdot)$ would be zero causing the bias also to be zero. The formula (13) for the optimal bandwidth leads in this case to

$$b_{opt}(u_0) = \left[\frac{2v_K}{d_K^2}\right]^{1/5}\left[\frac{\alpha_0(u_0)}{\alpha_0''(u_0)}\right]^{2/5} T^{-1/5}$$

leading to a large bandwidth if $\alpha_0''(u_0)$ is small and vice versa. As in Remark 2, this can be "translated" to the nonrescaled case.

Example 6 (tvGARCH processes). A tvGARCH(p,q) process satisfies the following representation

$$X_{t,T} = \sigma_{t,T}\,Z_t$$

where

$$\sigma^2_{t,T} = \alpha_0\left(\frac{t}{T}\right) + \sum_{j=1}^{p}\alpha_j\left(\frac{t}{T}\right)X^2_{t-j,T} + \sum_{i=1}^{q}\beta_i\left(\frac{t}{T}\right)\sigma^2_{t-i,T}, \quad (56)$$

where $\{Z_t\}$ are i.i.d. random variables with $\mathbf{E}Z_t = 0$ and $\mathbf{E}Z_t^2 = 1$. The corresponding stationary approximation at time u_0 is given by

$$\tilde{X}_t(u_0) = \sigma_t(u_0)\,Z_t \quad \text{for } t \in \mathbf{Z}$$

where

$$\sigma_t(u_0)^2 = \alpha_0(u_0) + \sum_{j=1}^{p}\alpha_j(u_0)\,\tilde{X}_{t-j}(u_0)^2 + \sum_{i=1}^{q}\beta_i(u_0)\,\sigma_{t-i}(u_0)^2. \quad (57)$$

Under the condition that $\sup_u \left(\sum_{j=1}^p \alpha_j(u) + \sum_{i=1}^q \beta_i(u) \right) < 1$, Subba Rao (2006), Section 5, has shown that $X_{t,T}^2 = \tilde{X}_t(u_0)^2 + O_p(|\frac{t}{T} - u_0| + \frac{1}{T})$. To obtain estimators of the parameters $\{\alpha_j(\cdot)\}$ and $\{\beta_i(\cdot)\}$, an approximation of the conditional quasi-likelihood is used, which is constructed as if the innovations $\{Z_t\}$ were Gaussian. As the infinite past is unobserved, an observable approximation of the conditional quasi-likelihood is

$$\ell_{t,T}(\boldsymbol{\theta}) = \frac{1}{2} \log w_{t,T}(\boldsymbol{\theta}) + \frac{X_{t,T}^2}{2\, w_{t,T}(\boldsymbol{\theta})} \text{ with } w_{t,T}(\boldsymbol{\theta}) = c_0(\boldsymbol{\theta}) + \sum_{j=1}^{t-1} c_j(\boldsymbol{\theta}) X_{t-j,T}^2, \quad (58)$$

where a recursive formula for $c_j(\boldsymbol{\theta})$ in terms of the parameters of interest, $\{\alpha_j\}$ and $\{\beta_i\}$, can be found in Berkes et al. (2003). Given that the derivatives of the time varying GARCH parameters exist, we can formally differentiate (57) to obtain

$$\frac{\partial \tilde{X}_t(u)^2}{\partial u} = \frac{\partial \sigma_t(u)^2}{\partial u} Z_t^2$$

$$\frac{\partial \sigma_t(u)^2}{\partial u} = \alpha_0'(u) + \sum_{j=1}^p \left(\alpha_j'(u)\, \tilde{X}_{t-j}(u)^2 + \alpha_j(u)\, \frac{\partial \tilde{X}_{t-j}(u)^2}{\partial u} \right)$$

$$+ \sum_{i=1}^q \left(\beta_i'(u)\, \sigma_{t-i}(u)^2 + \beta_i(u)\, \frac{\partial \sigma_{t-i}(u)^2}{\partial u} \right).$$

Subba Rao (2006) has shown that one can represent the above as a state-space representation which almost surely has a unique solution which is the derivative of $\tilde{X}_t(u)^2$ with respect to u. Thus, $X_{t,T}^2$ satisfies the expansion in (52). Moreover, Fryzlewicz and Subba Rao (2011) show geometric α-mixing of the tvGARCH process. Using these results and under some technical assumptions, it can be shown that Theorem 2 (i) and Theorem 3 hold for the local approximate conditional quasi-likelihood estimator. In particular, a result analogous to (55) holds true, where

$$\mathcal{L}(u, \boldsymbol{\theta}) = \mathbf{E}\Big(\log \big(c_0(\boldsymbol{\theta}) + \sum_{j=1}^{\infty} c_j(\boldsymbol{\theta}) \tilde{X}_{t-j}(u) \big) \Big) + \mathbf{E}\Big(\frac{\tilde{X}_t(u)^2}{c_0(\boldsymbol{\theta}) + \sum_{j=1}^{\infty} c_j(\boldsymbol{\theta}) \tilde{X}_{t-j}(u)^2} \Big).$$

Amado and Teräsvirta (2011) investigate parametric tvGARCH models, where the time varying parameters are modeled with the logistic transition function – see Section 2.6.

Similar methods as described in this section have also been applied in Koo and Linton (2010) who investigate semiparametric estimation of locally stationary diffusion models. They also prove a central limit theorem with a bias term as in (42). In their proofs, they use the stationary approximation $\tilde{X}_t(u_0)$ and the Taylor-type expansion (51). Vogt (2011) investigates nonlinear nonparametric models allowing for locally stationary regressors and a regression function that changes smoothly over time.

4. A general definition, linear processes and time varying spectral densities

The intuitive idea for a general definition is to require that locally around each rescaled time point U_0 the process $\{X_{t,T}\}$ can be approximated by a stationary process $\{\hat{X}_t(u_0)\}$ in a stochastic sense by using the property (4) (cf. Dahlhaus and Subba Rao, 2006). Vogt (2011) has formalized this by requiring that for each u_0 there exists a stationary process $\hat{X}_t(u_0)$ with

$$\|X_{t,T} - \hat{X}_t(u_0)| \leq \left(\left| \frac{t}{T} - u_0 \right| + \frac{1}{T} \right) U_{t,T}(u_0) \tag{59}$$

where $U_{t,T}(u_0)$ is a positive stochastic process fulfilling some uniform moment conditions. However up to now no general theory exists based on such a general definition. In the following we now move on toward a general theory for linear locally stationary processes. In some cases, we even assume Gaussianity or use Gaussian likelihood methods and their approximation. In this situation, a fairly general theory can be derived in which parametric and nonparametric inference problems, goodness of fit tests, bootstrap procedures, etc, can be treated in high generality. We use a general definition tailored for linear processes which implies (59).

4.1. Definition of linear locally stationary processes

We give this definition in terms of the time varying MA(∞) representation

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \sum_{j=-\infty}^{\infty} a_{t,T}(j)\,\varepsilon_{t-j}, \quad \text{where} \quad a_{t,T}(j) \approx a\left(\frac{t}{T}, j\right)$$

with coefficient functions $a(\cdot, j)$, which need to fulfill additional regularity function (dependent on the result to be proved – details are provided below). In several papers of the author, instead the time varying spectral representation

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \exp(i\lambda t)\, A_{t,T}(\lambda)\, d\xi(\lambda), \quad \text{where} \quad A_{t,T}(\lambda) \approx A\left(\frac{t}{T}, \lambda\right) \tag{60}$$

with the time varying transfer function $A(\cdot, \lambda)$ was used. Both representations are basically equivalent – see the derivation of (78). In the results presented below, we will always use the formulation "Under suitable regularity conditions ..." and refer the reader to the original paper. We conjecture, however, that all results can be reproved under Assumption 1. We emphasize that this is not an easy task, since in most situations, it means to transfer the proof from the frequency to the time domain. In that case, it would be worthwhile to require only martingale differences ε_t, since also some nonlinear processes admit such a representation.

Let

$$V(g) = \sup\Big\{ \sum_{k=1}^{m} \big|g(x_k) - g(x_{k-1})\big| : 0 \le x_o < \ldots < x_m \le 1,\ m \in \mathbf{N}\Big\} \quad (61)$$

be the total variation of g.

ASSUMPTION 1. *The sequence of stochastic processes $X_{t,T}$ has a representation*

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \sum_{j=-\infty}^{\infty} a_{t,T}(j)\,\varepsilon_{t-j} \quad (62)$$

where μ is of bounded variation and the ε_t are i.i.d. with $E\varepsilon_t = 0$, $E\varepsilon_s\varepsilon_t = 0$ for $s \neq t$, $E\varepsilon_t^2 = 1$. Let

$$\ell(j) := \begin{cases} 1, & |j| \le 1 \\ |j| \log^{1+\kappa} |j|, & |j| > 1 \end{cases}$$

for some $\kappa > 0$ and

$$\sup_t \big|a_{t,T}(j)\big| \le \frac{K}{\ell(j)} \quad \textit{(with } K \textit{ indep. of } T\textit{).} \quad (63)$$

Furthermore, we assume that there exist functions $a(\cdot, j) : (0,1] \to \mathbf{R}$ with

$$\sup_u \big|a(u,j)\big| \le \frac{K}{\ell(j)}, \quad (64)$$

$$\sup_j \sum_{t=1}^{T} \Big|a_{t,T}(j) - a\left(\frac{t}{T}, j\right)\Big| \le K, \quad (65)$$

$$V\,(a(\cdot, j)) \le \frac{K}{\ell(j)}. \quad (66)$$

The above assumptions are weak in the sense that only bounded variation is required for the coefficient functions. In particular for local results, stronger smoothness assumptions have to be imposed – for example in addition for some i

$$\sup_u \Big|\frac{\partial^i \mu(u)}{\partial u^i}\Big| \le K, \quad (67)$$

$$\sup_u \Big|\frac{\partial^i a(u,j)}{\partial u^i}\Big| \le \frac{K}{\ell(j)} \quad \text{for } j = 0, 1, \ldots \quad (68)$$

and instead of (65) the stronger assumption

$$\sup_{t,T} \Big|a_{t,T}(j) - a\left(\frac{t}{T}, j\right)\Big| \le \frac{K}{T\,\ell(j)}. \quad (69)$$

The construction with $a_{t,T}(j)$ and $a(t/T, j)$ looks complicated at first glance. The function $a(\cdot, j)$ is needed for rescaling and to impose necessary smoothness conditions, while the additional use of $a_{t,T}(j)$ makes the class rich enough to cover interesting cases such as tvAR-models (the reason for this in the AR(1) case can be understood from (49)). Cardinali and Nason (2010) created the term *close pair* for $\big(a(t/T, j), a_{t,T}(j)\big)$. Usually, additional moment conditions on ε_t are required.

It is straightforward to construct the stationary approximation and the derivative processes. We have

$$\tilde{X}_t(u) := \mu(u) + \sum_{j=-\infty}^{\infty} a(u, j)\, \varepsilon_{t-j}$$

and

$$\frac{\partial^i \tilde{X}_t(u)}{\partial u^i} = \frac{\partial^i \mu(u)}{\partial u^i} + \sum_{j=-\infty}^{\infty} \frac{\partial^i a(u, j)}{\partial u^i}\, \varepsilon_{t-j}$$

and it is easy to prove (59) and more general the expansion (51). We define the time varying spectral density by

$$f(u, \lambda) := \frac{1}{2\pi} \big|A(u, \lambda)\big|^2 \tag{70}$$

where

$$A(u, \lambda) := \sum_{j=-\infty}^{\infty} a(u, j)\, \exp(-i\lambda j), \tag{71}$$

and the time varying covariance of lag k at rescaled time u by

$$c(u, k) := \int_{-\pi}^{\pi} f(u, \lambda)\, \exp(i\lambda k)\, d\lambda = \sum_{j=-\infty}^{\infty} a(u, k + j)\, a(u, j). \tag{72}$$

$f(u, \lambda)$ and $c(u, k)$ are the spectral density and the covariance function of the stationary approximation $\tilde{X}_t(u)$. Under Assumption 1 and (69), it can be shown that

$$\operatorname{cov}\big(X_{[uT],T}, X_{[uT]+k,T}\big) = c(u, k) + O(T^{-1}) \tag{73}$$

uniformly in u and k – therefore we call $c(u, k)$ also the time varying covariance of the processes $X_{t,T}$. In Theorem 4, we show that $f(u, \lambda)$ is the uniquely defined time varying spectral density of $X_{t,T}$.

Example 7. (i) A simple example of a process $X_{t,T}$ which fulfills the above assumptions is $X_{t,T} = \mu(\frac{t}{T}) + \phi(\frac{t}{T}) Y_t$, where $Y_t = \Sigma_j\, a(j)\, \varepsilon_{t-j}$ is stationary with $|a(j)| \le K/\ell(j)$ and μ and ϕ are of bounded variation. If Y_t is an $AR(2)$ process with complex roots close to the unit circle, then Y_t shows a periodic behavior and

$\phi(\cdot)$ may be regarded as a time varying amplitude modulating function of the process $X_{t,T}$. $\phi(\cdot)$ may either be parametric or nonparametric.

(ii) The tvARMA(p,q) process

$$\sum_{j=0}^{p} \alpha_j\left(\frac{t}{T}\right) X_{t-j,T} = \sum_{k=0}^{q} \beta_k\left(\frac{t}{T}\right) \sigma\left(\frac{t-k}{T}\right) \varepsilon_{t-k} \tag{74}$$

where ε_t are i.i.d. with $E\varepsilon_t = 0$ and $E\varepsilon_t^2 < \infty$ and all $\alpha_j(\cdot)$, $\beta_k(\cdot)$ and $\sigma^2(\cdot)$ are of bounded variation with $\alpha_0(\cdot) \equiv \beta_0(\cdot) \equiv 1$ and $\sum_{j=0}^{p} \alpha_j(u) z^j \neq 0$ for all u and all $|z| \leq 1 + \delta$ for some $\delta > 0$, fulfills Assumption 1. If the parameters are differentiable with bounded derivatives, then also (67)–(69) are fulfilled (for i = 1). The time varying spectral density is

$$f(u,\lambda) = \frac{\sigma^2(u)}{2\pi} \frac{|\sum_{k=0}^{q} \beta_k(u) \exp(i\lambda k)|^2}{|\sum_{j=0}^{p} \alpha_j(u) \exp(i\lambda j)|^2}. \tag{75}$$

This is proved by Dahlhaus and Polonik (2006). $\alpha_j(\cdot)$ and $\beta_k(\cdot)$ may either be parametric or nonparametric.

The time varying MA(∞) representation (62) can easily be transformed into a time varying spectral representation as used, e.g., in the study by Dahlhaus (1997, 2000). If the ε_t are assumed to be stationary, then there exists a Cramér representation (cf. Brillinger, 1981)

$$\varepsilon_t = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \exp(i\lambda t)\, d\xi(\lambda) \tag{76}$$

where $\xi(\lambda)$ is a process with mean 0 and orthonormal increments. Let

$$A_{t,T}(\lambda) := \sum_{j=-\infty}^{\infty} a_{t,T}(j)\, \exp(-i\lambda j). \tag{77}$$

Then

$$X_{t,T} = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \exp(i\lambda t)\, A_{t,T}(\lambda)\, d\xi(\lambda). \tag{78}$$

(69) now implies

$$\sup_{t,\lambda} \left| A_{t,T}(\lambda) - A\left(\frac{t}{T}, \lambda\right) \right| \leq K T^{-1} \tag{79}$$

which was assumed in the above cited papers. Conversely, if we start with (78) and (79), then we can conclude from adequate smoothness conditions on $A(u,\lambda)$ to the conditions of Assumption 1.

We now state a uniqueness property of our spectral representation. The Wigner-Ville spectrum for fixed T (cf. Martin and Flandrin, 1985) is

$$f_T(u,\lambda) := \frac{1}{2\pi} \sum_{s=-\infty}^{\infty} \operatorname{cov}(X_{[uT-s/2],T}, X_{[uT+s/2],T}) \exp(-i\lambda s)$$

with $X_{t,T}$ as in (62) (either with the coefficient extended as constants for $u \notin [0,1]$ or set to 0). Below we prove that $f_T(u,\lambda)$ tends in squared mean to $f(u,\lambda)$ as defined in (70). Therefore, it is justified to call $f(u,\lambda)$ the time varying spectral density of the process.

THEOREM 4. *If $X_{t,T}$ is locally stationary and fulfills Assumption 1 and (68) for all j, then we have for all $u \in (0,1)$*

$$\int_{-\pi}^{\pi} \left| f_T(u,\lambda) - f(u,\lambda) \right|^2 d\lambda = o(1).$$

PROOF. The result was proved by Dahlhaus (1996b) under a different set of conditions. It is not very difficult to prove the result also under the present conditions. □

As a consequence, the time varying spectral density $f(u,\lambda)$ is uniquely defined. If in addition, the process $X_{t,T}$ is non-Gaussian, then even $A(u,\lambda)$ and, therefore, also the coefficients $a(u,j)$ are uniquely determined which may be proved similarly by considering higher-order spectra. Since $\mu(t/T)$ is the mean of the process, it is also uniquely determined. This is remarkable, since in the nonrescaled case, time varying processes do not have a unique spectral density or a unique time varying spectral representation (cf. Priestley, 1981, Chapter 11.1; Mélard and Herteleer-de-Schutter, 1989). $f(u,\lambda)$ from Theorem 4 has been called instantaneous spectrum (in particular for tvAR process – cf. Kitagawa and Gersch, 1985). The above theorem gives a theoretical justification for this definition.

There is a huge benefit from having a unique time varying spectral density. We now give an example for this. We derive the limit of the Kullback-Leibler information for Gaussian processes and show that it depends on $f(u,\lambda)$. Replacing this by a spectral estimate will lead to a quasi-likelihood for parametric models similar to the Whittle likelihood for stationary processes. Without a unique spectral density such a construction were not possible.

Consider the exact Gaussian maximum likelihood estimate

$$\hat{\eta}_T^{ML} := \operatorname*{argmin}_{\eta \in \Theta_\eta} \mathcal{L}_T^E(\eta)$$

where η is a finite-dimensional parameter (as in (20)) and

$$\mathcal{L}_T^E(\eta) = \frac{1}{2}\log(2\pi) + \frac{1}{2T}\log\det\Sigma_\eta + \frac{1}{2T}(\mathbf{X}-\mu_\eta)'\Sigma_\eta^{-1}(\mathbf{X}-\mu_\eta) \qquad (80)$$

with $\mathbf{X} = (X_{1,T}, \ldots, X_{T,T})'$, $\mu_\eta = \big(\mu_\eta(1/T), \ldots, \mu_\eta(T/T)\big)'$, and Σ_η being the covariance matrix of the model. Under certain regularity conditions $\hat{\eta}_T^{ML}$ will converge to

$$\eta_0 := \operatorname*{argmin}_{\eta \in \Theta_\eta} \mathcal{L}(\eta) \tag{81}$$

where

$$\mathcal{L}(\eta) := \lim_{T\to\infty} \mathbf{E}\, \mathcal{L}_T^E(\eta).$$

If the model is correct, then typically η_0 is the true parameter value. Otherwise, it is some "projection" onto the parameter space. It is, therefore, important to calculate $\mathcal{L}(\eta)$, which is equivalent to the calculation of the Kullback-Leibler information divergence.

THEOREM 5. *Let $X_{t,T}$ be a locally stationary process with true mean and spectral density curves $\mu(\cdot)$, $f(u,\lambda)$ and model curves $\mu_\eta(\cdot)$, $f_\eta(u,\lambda)$, respectively. Under suitable regularity conditions, we have*

$$\begin{aligned}\mathcal{L}(\eta) &= \lim_{T\to\infty} \mathbf{E}\, \mathcal{L}_T^E(\eta) \\ &= \frac{1}{4\pi} \int_0^1 \int_{-\pi}^{\pi} \left\{ \log 4\pi^2 f_\eta(u,\lambda) + \frac{f(u,\lambda)}{f_\eta(u,\lambda)} \right\} d\lambda\, du + \frac{1}{4\pi} \int_0^1 \frac{\big(\mu_\eta(u) - \mu(u)\big)^2}{f_\eta(u,0)}\, du.\end{aligned}$$

PROOF. See Dahlhaus (1996b), Theorem 3.4 . □

The Kullback-Leibler information divergence for stationary processes is obtained from this as a special case (cf. Parzen, 1983).

Example 8. Suppose that the model is stationary, i.e., $f_\eta(\lambda) := f_\eta(u,\lambda)$ and $m := \mu_\eta(u)$ do not depend on u. Then,

$$\mathcal{L}(\eta) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left\{ \log 4\pi^2 f_\eta(\lambda) + \frac{\int_0^1 f(u,\lambda)\, du}{f_\eta(\lambda)} \right\} d\lambda + \frac{1}{4\pi}\, f_\eta(0)^{-1} \int_0^1 \big(m - \mu(u)\big)^2 du$$

i.e., $m_0 = \int_0^1 \mu(u) du$ and $f_{\eta_0}(\lambda)$ give the best approximation to the time integrated true spectrum $\int_0^1 f(u,\lambda)\, du$. These are the values which are "estimated" by the MLE or a quasi-MLE, if a stationary model is fitted to locally stationary data.

Given the form of $\mathcal{L}(\eta)$ as in Theorem 5, we can now suggest a quasi-likelihood criterion

$$\mathcal{L}_T^{QL}(\eta) = \frac{1}{4\pi} \int_0^1 \int_{-\pi}^{\pi} \Big\{ \log 4\pi^2 f_\eta(u,\lambda) + \frac{\hat{f}(u,\lambda)}{f_\eta(u,\lambda)} \Big\} d\lambda\, du + \frac{1}{4\pi} \int_0^1 \frac{\big(\mu_\eta(u) - \hat{\mu}(u)\big)^2}{f_\eta(u,0)}\, du$$

where $\hat{f}(u,\lambda)$ and $\hat{\mu}(u)$ are suitable nonparametric estimates of $f(u,\lambda)$ and $\mu(u)$, respectively. The block Whittle likelihood $\mathcal{L}_T^{BW}(\eta)$ in (21) and the generalized Whittle likelihood $\mathcal{L}_T^{GW}(\eta)$ in (89) are of this form.

We now calculate the Fisher information matrix

$$\Gamma := \lim_{T\to\infty} T\, \mathbf{E}_{\eta_0} \big(\nabla \mathcal{L}_T^E(\eta_0)\big) \big(\nabla \mathcal{L}_T^E(\eta_0)\big)'$$

in order to study efficiency of parameter estimates (see also Theorem 8).

THEOREM 6. *Let $X_{t,T}$ be a locally stationary process with correctly specified mean curve $\mu_\eta(u)$ and time varying spectral density $f_\eta(u,\lambda)$. Under suitable regularity conditions, we have*

$$\Gamma = \frac{1}{4\pi} \int_0^1 \int_{-\pi}^{\pi} \big(\nabla \log f_{\eta_0}\big)\big(\nabla \log f_{\eta_0}\big)'\, d\lambda\, du + \frac{1}{2\pi} \int_0^1 \big(\nabla \mu_{\eta_0}(u)\big)\big(\nabla \mu_{\eta_0}(u)\big)' f_{\eta_0}^{-1}(u,0)\, du.$$

PROOF. See Dahlhaus (1996b), Theorem 3.6. □

We now briefly discuss how the time varying spectral density can be estimated. Following the discussion in the last section, we start with a classical "stationary" smoothed periodogram estimate on a segment. Let $I_T(u,\lambda)$ be the tapered periodogram on a segment of length N about u as defined in (19). Even in the stationary case, $I_T(u,\lambda)$ is not a consistent estimate of the spectrum and we have to smooth it over neighboring frequencies. Let, therefore,

$$\hat{f}_T(u,\lambda) := \frac{1}{b_f} \int K_f\left(\frac{\lambda-\mu}{b_f}\right) I_T(u,\mu)\, d\mu \tag{82}$$

where K_f is a symmetric kernel with $\int K_f(x)\, dx = 1$ and b_f is the bandwidth in frequency direction. Theorem 5.5 below shows that the estimate is implicitly also a kernel

estimate in time direction with kernel

$$K_t(x) := \left\{ \int_0^1 h(x)^2 dx \right\}^{-1} h(x+1/2)^2, \qquad x \in [-1/2, 1/2] \tag{83}$$

and bandwidth $b_t := N/T$, that is, the estimate behaves like a kernel estimates with two convolution kernels in frequency and time direction. We mention that an asymptotically equivalent estimate is the kernel estimate

$$\tilde{f}_T(u,\lambda) := \frac{2\pi}{T^2} \sum_{t=1}^{T} \sum_{j=1}^{T} \int \frac{1}{b_t} K_t\left(\frac{u - t/T}{b_t}\right) \frac{1}{b_f} K_f\Big(\frac{\lambda - \lambda_j}{b_f}\Big)\, J_T\Big(\frac{t}{T}, \lambda_j\Big) \tag{84}$$

with the preperiodogram $J_T(u,\lambda)$ as defined in (88). One may also replace the integral in frequency direction in (82) by a sum over the Fourier frequencies.

THEOREM 7. *Let $X_{t,T}$ be a locally stationary process with $\mu(\cdot) \equiv 0$. Under suitable regularity conditions, we have*

$$(i)\ \ \mathbf{E} I_T(u,\lambda) = f(u,\lambda) + \frac{1}{2} b_t^2 \int_{-1/2}^{1/2} x^2 K_t(x)\, dx\, \frac{\partial^2}{\partial u^2} f(u,\lambda) + o(b_t^2) + O\left(\frac{\log(b_t T)}{b_t T}\right);$$

$$(ii)\ \mathbf{E}\hat{f}_T(u,\lambda) = f(u,\lambda) + \frac{1}{2} b_t^2 \int_{-1/2}^{1/2} x^2 K_t(x)\, dx\, \frac{\partial^2}{\partial u^2} f(u,\lambda) + \frac{1}{2} b_f^2 \int_{-1/2}^{1/2} x^2 K_f(x)\, dx\, \frac{\partial^2}{\partial \lambda^2} f(u,\lambda) + o\Big(b_t^2 + b_f^2 + \frac{\log(b_t T)}{b_t T}\Big);$$

$$(iii)\ \operatorname{var}\big(\hat{f}_T(u,\lambda)\big) = \big(b_t b_f T\big)^{-1} 2\pi f(u,\lambda)^2 \int_{-1/2}^{1/2} K_t(x)^2 dx \int_{-1/2}^{1/2} K_f(x)^2 dx\, \big(1 + \delta_{\lambda 0}\big).$$

PROOF. A sketch of the proof can be found in Dahlhaus (1996c), Theorem 2.2. □

Note, that the first-bias term of $\hat{f}$ is due to nonstationarity, while the second is due to the variation of the spectrum in frequency direction.

As in Remark 1, one may now minimize the relative mean squared error $\mathrm{RMSE}(\hat{f}) := E\big(\hat{f}(u,\lambda)/f(u,\lambda) - 1\big)^2$ with respect to b_f, b_t (i.e., N), K_f and K_t (i.e., the data

taper h). This has been done by Dahlhaus (1996c), Theorem 2.3. The result says that with

$$\Delta_u := \frac{\partial^2}{\partial u^2} f(u,\lambda) \Big/ f(u,\lambda) \quad \text{and} \quad \Delta_\lambda := \frac{\partial^2}{\partial \lambda^2} f(u,\lambda) \Big/ f(u,\lambda),$$

the optimal RMSE is obtained with

$$b_t^{\text{opt}} = T^{-1/6}(576\pi)^{1/6} \left(\frac{\Delta_\lambda}{\Delta_u^5} \right)^{1/12}, \quad b_f^{\text{opt}} = T^{-1/6}(576\pi)^{1/6} \left(\frac{\Delta_u}{\Delta_\lambda^5} \right)^{1/12}$$

and optimal kernels $K_t^{\text{opt}}(x) = K_f^{\text{opt}}(x) = 6\left(1/4 - x^2\right)$ with optimal rate $T^{-2/3}$.

The relations $b_t = N/T$ and (83) immediately lead to the optimal segment length and the optimal data taper h. The result of Theorem 5.5 is quite reasonable: If the degree of nonstationarity is small, then Δ_u is small and b_t^{opt} gets large. If the variation of f is small in frequency direction, then Δ_λ is small and b_t^{opt} gets smaller (more smoothing is put in frequency direction than in time direction). This is another example, how the bias due to nonstationarity can be quantified with the approach of local stationarity and balanced with another bias term and a variance term. Of course, the data-adaptive choice of the bandwidth parameters remains to be solved. Asymptotic normality of the estimates can be derived from Theorem 11 (cf. Dahlhaus, 2009, Example 4.2).

Rosen et al. (2009) estimate the logarithm of the local spectrum by using a Bayesian mixture of splines. They assume that the log spectrum on a partition of the data is a mixture of individual log spectra and use a mixture of smoothing splines with time varying mixing weights to estimate the evolutionary log spectrum. Guo et al. (2003) use a smoothing spline ANOVA to estimate the time varying log spectrum.

5. Gaussian likelihood theory for locally stationary processes

The basics of the likelihood theory for univariate stationary processes were laid by Whittle (1953, 1954). His work was much later taken up and continued by many others. Among the large number of papers, we mention the results by Dzhaparidze (1971) and Hannan (1973) for univariate time series, Dunsmuir (1979) for multivariate time series and, e.g., Hosoya and Taniguchi (1982) for misspecified multivariate time series. A general overview over this likelihood theory and, in particular, Whittle estimates for stationary models may be found in the monographs Dzhaparidze (1986) and Taniguchi and Kakizawa (2000).

From a practical point of view, the most famous outcome of this theory is the Whittle likelihood

$$\frac{1}{4\pi} \int_{-\pi}^{\pi} \left\{ \log 4\pi^2 f_\eta(\lambda) + \frac{I_T(\lambda)}{f_\eta(\lambda)} \right\} d\lambda \tag{85}$$

as an approximation of the negative log Gaussian likelihood (80), where $I_T(\lambda)$ is the periodogram. This likelihood has been used also beyond the classical framework – for example by Mikosch et al. (1995) for linear processes where the innovations have

heavy tailed distributions, by Fox and Taqqu (1986) for long-range dependent processes, and by Robinson (1995) to construct semiparametric estimates for long-range dependent processes.

The outcome of this likelihood theory goes far beyond the construction of the Whittle likelihood. Its technical core is the theory of Toeplitz matrices and, in particular, the approximation of the inverse of a Toeplitz matrix by the Toeplitz matrix of the inverse function. It is essentially this approximation which leads from the ordinary Gaussian likelihood to the Whittle likelihood. Beyond that, the theory can be used to derive the convergence of experiments for Gaussian stationary processes in the Hájek-Le Cam sense, construct the properties of many tests, and derive the properties of the exact MLE and the Whittle estimate (cf. Dzhaparidze, 1986; Taniguchi and Kakizawa, 2000).

For locally stationary processes, it turns out that this likelihood theory can be generalized in a nice way such that the classical likelihood theory for stationary processes arises as a special case. Technically speaking, this is achieved by a generalization of Toeplitz matrices tailored especially for locally stationary processes (the matrix $U_T(\phi)$ defined in (92)).

Some results coming from this theory have already been stated in Section 4, namely the limit of the Kullback-Leibler information divergence in Theorem 5 and the limit of the Fischer information in Theorem 6. We now describe further results. We start with a decomposition of the periodogram leading to a Whittle-type likelihood. We have

$$\begin{aligned} I_T(\lambda) &= \frac{1}{2\pi T}\Big|\sum_{r=1}^{T} X_r \exp(-i\lambda r)\Big|^2 \\ &= \frac{1}{2\pi}\sum_{k=-(T-1)}^{T-1}\left(\frac{1}{T}\sum_{t=1}^{T-|k|} X_t X_{t+|k|}\right)\exp(-i\lambda k) \qquad (86) \\ &= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2\pi}\sum_{\substack{k \\ 1\le[t+0.5+k/2],[t+0.5-k/2]\le T}} X_{[t+0.5+k/2],T}\, X_{[t+0.5-k/2],T}\exp(-i\lambda k) \\ &= \frac{1}{T}\sum_{t=1}^{T} J_T\left(\frac{t}{T},\lambda\right), \qquad (87) \end{aligned}$$

where the so-called preperiodogram

$$J_T(u,\lambda) := \frac{1}{2\pi}\sum_{\substack{k \\ 1\le[uT+0.5+k/2],[uT+0.5-k/2]\le T}} X_{[uT+0.5+k/2],T}\, X_{[uT+0.5-k/2],T}\ \exp(-i\lambda k) \qquad (88)$$

may be regarded as a local version of the periodogram at time t. While the ordinary periodogram $I_T(\lambda)$ is the Fourier transform of the covariance estimator of lag k over the whole segment (see (86)), the preperiodogram just uses the pair $X_{[t+0.5+k/2]}X_{[t+0.5-k/2]}$ as a kind of "local estimator" of the covariance of lag k at time t (note that $[t+0.5+$

$k/2] - [t + 0.5 - k/2] = k\big)$. The preperiodogram was introduced by Neumann and von Sachs (1997) as a starting point for a wavelet estimate of the time varying spectral density. The above decomposition means that the periodogram is the average of the preperiodogram over time.

If we replace $I_T(\lambda)$ in (85) by the above average of the preperiodogram and afterward replace the model spectral density $f_\eta(\lambda)$ by the time varying spectral density $f_\eta(u,\lambda)$ of a nonstationary model, we obtain the *generalized Whittle likelihood*

$$\mathcal{L}_T^{GW}(\eta) := \frac{1}{T}\sum_{t=1}^{T}\frac{1}{4\pi}\int_{-\pi}^{\pi}\left\{\log 4\pi^2 f_\eta\Big(\frac{t}{T},\lambda\Big) + \frac{J_T(\frac{t}{T},\lambda)}{f_\eta(\frac{t}{T},\lambda)}\right\}d\lambda. \tag{89}$$

If the fitted model is stationary, i.e., $f_\eta(u,\lambda) = f_\eta(\lambda)$, then (due to (87)) the above likelihood is identical to the Whittle likelihood and we obtain the classical Whittle estimator. Thus, the above likelihood is a true generalization of the Whittle likelihood to nonstationary processes. In Theorem 9, we show that this likelihood is a very close approximation to the Gaussian log-likelihood for locally stationary processes. In particular (we conjecture that), it is a better approximation than the block Whittle likelihood $\mathcal{L}_T^{BW}(\eta)$ from (21).

We now briefly state the asymptotic normality result in the parametric case. An example is the tvAR(2) model with polynomial parameter curves from Section 2.4. Let

$$\hat{\eta}_T^{GW} := \operatorname*{argmin}_{\eta\in\Theta_\eta}\mathcal{L}_T^{GW}(\eta) \tag{90}$$

be the corresponding quasi-likelihood estimate, $\hat{\eta}_T^{ML}$ be the Gaussian MLE defined in (80), and η_0 as in (81), i.e., the model may be misspecified.

THEOREM 8. *Let $X_{t,T}$ be a locally stationary process. Under suitable regularity conditions we have in the case $\mu(\cdot) = \mu_\eta(\cdot) = 0$*

$$\sqrt{T}\big(\hat{\eta}_T^{GW} - \eta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \Gamma^{-1}V\Gamma^{-1}\big) \quad \textit{and} \quad \sqrt{T}\big(\hat{\eta}_T^{ML} - \eta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \Gamma^{-1}V\Gamma^{-1}\big)$$

with

$$\Gamma_{ij} = \frac{1}{4\pi}\int_0^1\int_{-\pi}^{\pi}\big(f - f_{\eta_0}\big)\nabla_{ij}f_{\eta_0}^{-1}\,d\lambda\,du + \frac{1}{4\pi}\int_0^1\int_{-\pi}^{\pi}\big(\nabla_i \log f_{\eta_0}\big)\big(\nabla_j \log f_{\eta_0}\big)\,d\lambda\,du$$

and

$$V_{ij} = \frac{1}{4\pi}\int_0^1\int_{-\pi}^{\pi} f\left(\nabla_i f_\eta^{-1}\right) f\left(\nabla_j f_\eta^{-1}\right)d\lambda\,du.$$

If the model is correctly specified, then $V = \Gamma$ and Γ is the same as in Theorem 6 – that is both estimates are asymptotically Fisher efficient. Even more the sequence

of experiments is locally asymptotically normal (LAN) and both estimates are locally asymptotically minimax.

PROOF. See Dahlhaus (2000), Theorem 3.1. LAN and LAM have been proved for the MLE in Dahlhaus (1996b), Theorem 4.1 and 4.2, – these results together with the LAM property of the generalized Whittle estimate also follow from the technical lemmas by Dahlhaus (2000) (cf. Remark 3.3 in that paper). □

The corresponding result in the multivariate case and in the case $\mu(\cdot) \neq 0$ or $\mu_\eta(\cdot) \neq 0$ can be found in the study by Dahlhaus (2000), Theorem 3.1.

A deeper investigation of $\mathcal{L}_T^{GW}(\eta)$ reveals that it can be derived from the Gaussian log-likelihood by an approximation of the inverse of the covariance matrix. Let $\underline{X} = \left(X_{1,T}, \ldots, X_{T,T}\right)'$, $\underline{\mu} = \left(\mu(\frac{1}{T}), \ldots, \mu(\frac{T}{T})\right)'$, and $\Sigma_T(A,B)$ and $U_T(\phi)$ be $T \times T$ matrices with (r,s)-entry

$$\Sigma_T(A,B)_{r,s} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp\left(i\lambda(r-s)\right) A_{r,T}(\lambda)\, B_{s,T}(-\lambda) d\lambda \tag{91}$$

and

$$U_T(\phi)_{r,s} = \int_{-\pi}^{\pi} \exp\left(i\lambda(r-s)\right) \phi\Big(\frac{1}{T}\Big[\frac{r+s}{2}\Big]^*, \lambda\Big) d\lambda \tag{92}$$

$(r,s = 1, \ldots T)$, where the functions $A_{r,T}(\lambda)$, $B_{r,T}(\lambda)$, and $\phi(u,\lambda)$ fulfill certain regularity conditions ($A_{r,T}(\lambda)$ $B_{r,T}(\lambda)$ are transfer functions or their derivatives as defined in (77)). $[x]^* = [x]$ denotes the largest integer less or equal to x (we have added the * to discriminate the notation from brackets). Direct calculation shows that

$$\begin{aligned} \mathcal{L}_T^{GW}(\eta) = \frac{1}{4\pi} \frac{1}{T} \sum_{t=1}^{T} \int_{-\pi}^{\pi} \log\left[4\pi^2 f_\eta\left(\frac{t}{T}, \lambda\right)\right] d\lambda + \frac{1}{8\pi^2 T} \left(\underline{X} - \underline{\mu}_\eta\right)' \\ \times U_T\left(f_\eta^{-1}\right) \left(\underline{X} - \underline{\mu}_\eta\right). \end{aligned} \tag{93}$$

Furthermore, the exact Gaussian likelihood is

$$\mathcal{L}_T^E(\eta) := \frac{1}{2} \log(2\pi) + \frac{1}{2T} \log\det \Sigma_\eta + \frac{1}{2T} \left(\underline{X} - \underline{\mu}_\eta\right)' \Sigma_\eta^{-1} \left(\underline{X} - \underline{\mu}_\eta\right) \tag{94}$$

where $\Sigma_\eta = \Sigma_T(A_\eta, A_\eta)$.

Proposition 1 below states that $U_T(\frac{1}{4\pi^2} f_\eta^{-1})$ is an approximation of Σ_η^{-1}. Together with the generalization of the Szegö identity in Proposition 2, this implies that $\mathcal{L}_T^{GW}$ is an approximation of $\mathcal{L}_T^E$ (see Theorem 9). If the model is stationary, then A_η is constant in time and $\Sigma_\eta = \Sigma_T(A_\eta, A_\eta)$ is the Toeplitz matrix of the spectral density $f_\eta(\lambda) = \frac{1}{2\pi}|A_\eta|^2$, whereas $U_T(\frac{1}{4\pi^2} f_\eta^{-1})$ is the Toeplitz matrix of $\frac{1}{4\pi^2} f_\eta^{-1}$. This is the classical matrix-approximation leading to the Whittle likelihood (cf. Dzhaparidze, 1986).

PROPOSITION 1. *Under suitable regularity conditions we have for each* $\varepsilon > 0$ *for the Euclidean norm*

$$\frac{1}{T}\left\|\Sigma_T(A,A)^{-1} - U_T\left(\{2\pi A\bar{A}'\}^{-1}\right)\right\|_2^2 = O(T^{-1+\varepsilon}) \tag{95}$$

and

$$\frac{1}{T}\left\|U_T(\phi)^{-1} - U_T\left(\{4\pi^2\phi\}^{-1}\right)\right\|_2^2 = O(T^{-1+\varepsilon}).$$

PROOF. See Dahlhaus (2000), Proposition 2.4. □

By using the above approximation, it is possible to prove the following generalization of the Szegö identity (cf. Grenander and Szegö, 1958, Section 5.2) to locally stationary processes.

PROPOSITION 2. *Under suitable regularity conditions, we have with* $f(u,\lambda) = \frac{1}{2\pi}|A(u,\lambda)|^2$ *for each* $\varepsilon > 0$

$$\frac{1}{T}\log\det\Sigma_T(A,A) = \frac{1}{2\pi}\int_0^1\int_{-\pi}^{\pi}\log\left[2\pi f(u,\lambda)\right]d\lambda\,du + O(T^{-1+\varepsilon}).$$

If $A = A_\eta$ *depends on a parameter* η*, then the* $O(T^{-1+\varepsilon})$ *term is uniform in* η*.*

PROOF. See Dahlhaus (2000), Proposition 2.5. □

In certain situations, the right-hand side can be written in the form $\int_0^1 \log\left(2\pi\sigma^2(u)\right)du$, where $\sigma^2(u)$ is the one-step prediction error at time u.

The mathematical core of the above results consists of the derivation of properties of products of matrices $\Sigma_T(A,B)$, $\Sigma_T(A,A)^{-1}$, and $U_T(\phi)$. These properties are derived by Dahlhaus (2000) in Lemmas A.1, A.5, A.7, and A.8. These results are generalizations of corresponding results in the stationary case proved by several authors before.

We now state the properties of the different likelihoods.

THEOREM 9. *Under suitable regularity conditions, we have for* $k = 0, 1, 2$

(i) $\sup_{\eta\in\Theta_\eta}\left|\nabla^k\left\{\mathcal{L}_T^{GW}(\eta) - \mathcal{L}_T^{E}(\eta)\right\}\right| \xrightarrow{P} 0,$

(ii) $\sup_{\theta\in\Theta_\eta}\left|\nabla^k\left\{\mathcal{L}_T^{GW}(\eta) - \mathcal{L}(\eta)\right\}\right| \xrightarrow{P} 0,$

(iii) $\sup_{\eta\in\Theta_\eta}\left|\nabla^k\left\{\mathcal{L}_T^{E}(\eta) - \mathcal{L}(\eta)\right\}\right| \xrightarrow{P} 0.$

PROOF. See Dahlhaus (2000) Theorem 3.1. □

Under stronger assumptions one may also conclude that $\hat{\eta}_T^{GW} - \hat{\eta}_T^{ML} = O_p(T^{-1+\varepsilon})$, which means that $\hat{\eta}_T^{GW}$ is a close approximation of the MLE. A sketch of the proof is given in Dahlhaus (2000), Remark 3.4.

REMARK 5. It is interesting to compare the generalized Whittle estimate $\hat{\eta}_T^{GW}$ and its underlying approximation $U_T(\frac{1}{4\pi^2} f_\eta^{-1})$ of Σ_η^{-1} with the block Whittle estimate $\hat{\eta}_T^{BW}$ defined in (21). There some overlapping block Toeplitz matrices are used as an approximation which we regard as worse. A similar result as in Proposition 2 has been proved in Lemma 4.7 of Dahlhaus (1996a) for this approximation. We conjecture that also a similar result as in Theorem 9 with $\mathcal{L}_T^{BW}(\eta)$ can be proved and even more that $\hat{\eta}_T^{BW} - \hat{\eta}_T^{ML} = O_p\big(\frac{N}{T^{1-\varepsilon}} + \frac{1}{N}\big)$ (this is more a vague guess than a solid conjecture), which means that the latter approximation and presumably also the estimate $\hat{\eta}_T^{BW}$ are worse. It would be interesting to have more rigorous results and a careful simulation study with a comparison of both estimates. □

We now remember the generalized Whittle likelihood from (89), which was

$$\mathcal{L}_T^{GW}(\eta) = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{4\pi} \int_{-\pi}^{\pi} \left\{ \log 4\pi^2 f_\eta\Big(\frac{t}{T}, \lambda\Big) + \frac{J_T(\frac{t}{T}, \lambda)}{f_\eta(\frac{t}{T}, \lambda)} \right\} d\lambda.$$

Contrary to the true Gaussian likelihood, this is a sum over time and the summands can be interpreted as a local log-likelihood at time point t. We, therefore, define

$$\ell_{t,T}^*(\boldsymbol{\theta}) := \frac{1}{4\pi} \int_{-\pi}^{\pi} \left\{ \log 4\pi^2 f_{\boldsymbol{\theta}}(\lambda) + \frac{J_T(\frac{t}{T}, \lambda)}{f_{\boldsymbol{\theta}}(\lambda)} \right\} d\lambda. \tag{96}$$

(to avoid confusion we mention that we use the notation η for a finite-dimensional parameter which determines the whole curve, that is $\boldsymbol{\theta}(\cdot) = \boldsymbol{\theta}_\eta(\cdot)$ and $f_\eta(u, \lambda) = f_{\boldsymbol{\theta}_\eta(u)}(\lambda)$). We now can construct all nonparametric estimates (26)–(30) with $\ell_{t,T}(\boldsymbol{\theta})$ replaced by $\ell_{t,T}^*(\boldsymbol{\theta})$ leading in each of the five cases to an alternative local quasi-likelihood estimate.

The parametric estimator (30) with this local likelihood is the estimate $\hat{\eta}_T^{GW}$ from above. The orthogonal series estimator (28) with $\ell_{t,T}^*(\boldsymbol{\theta})$ has been investigated for a truncated wavelet series expansion together with nonlinear thresholding by Dahlhaus and Neumann (2001). The method is fully automatic and adapts to different smoothness classes. It is shown that the usual rates of convergence in Besov classes are attained up to a logarithmic factor. The nonparametric estimator (29) with $\ell_{t,T}^*(\boldsymbol{\theta})$ is studied by Dahlhaus and Polonik (2006). Rates of convergence, depending on the metric entropy of the function space, are derived. This includes in particular maximum likelihood estimates derived under shape restriction. The main tool for deriving these results is the so-called empirical spectral processes discussed in the next section. The kernel estimator (26) with $\ell_{t,T}^*(\boldsymbol{\theta})$ has been investigated in Dahlhaus (2009), Example 3.6. Uniform convergence has been proved by Dahlhaus and Polonik (2009), Section 4 (see also Example 9 and Theorem 13 below). The local polynomial fit (27) has not been investigated yet in combination with this likelihood.

The whole topic needs a more careful investigation – both theoretically and from a practical point including simulations and data-examples.

6. Empirical spectral processes

We now emphasize the relevance of the empirical spectral process for linear locally stationary time series. The theory of empirical processes not only plays a major role in proving theoretical results for statistical methods but also provides a deeper understanding of many techniques and the arising problems. The theory was first developed for stationary processes (cf. Dahlhaus, 1988; Mikosch and Norvaisa, 1997; Fay and Soulier, 2001) and then extended to locally stationary processes in Dahlhaus and Polonik (2006, 2009) and Dahlhaus (2009). The empirical spectral process is indexed by classes of functions. Basic results that later lead to several statistical applications are a functional central limit theorem, a maximal exponential inequality and a Glivenko-Cantelli type convergence result. All results use conditions based on the metric entropy of the index class. Many results stated earlier in this article have been proved by using these techniques.

The empirical spectral process is defined by

$$E_T(\phi) := \sqrt{T}\left(F_T(\phi) - F(\phi)\right)$$

where

$$F(\phi) := \int_0^1 \int_{-\pi}^{\pi} \phi(u,\lambda) f(u,\lambda) d\lambda\, du \tag{97}$$

is the generalized spectral measure and

$$F_T(\phi) := \frac{1}{T}\sum_{t=1}^{T} \int_{-\pi}^{\pi} \phi\left(\frac{t}{T},\lambda\right) J_T\left(\frac{t}{T},\lambda\right) d\lambda \tag{98}$$

the empirical spectral measure with the preperiodogram as defined in (88).

We first give an overview of statistics that can be written in the form $F_T(\phi)$ - several of them have already been discussed earlier in this article (K_T always denotes a kernel function).

1.	$\phi(u,\lambda) = K_T(u_0-u)\cos(\lambda k)$	local covariance estimator	(9) a.s.; Remark 9
2.	$\phi(u,\lambda) = K_T(u_0-u)\, K_T(\lambda_0-\lambda)$	spectral density estimator	(84) a.s.; Remark 9
3.	$\phi(u,\lambda) = K_T(u_0-u)\nabla f_{\boldsymbol{\theta}_0}(u,\lambda)^{-1}$	$\nabla \mathcal{L}_T^{GW}(u_0,\boldsymbol{\theta}_0)$, $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0(u_0)$	Example 9
4.	$\phi(u,\lambda) \approx K_T(u_0-u)\nabla f_{\boldsymbol{\theta}_0}(u,\lambda)^{-1}$	local least squares	Example 1; Remark 9
5.	$\phi(u,\lambda) = \nabla f_{\eta_0}(u,\lambda)^{-1}$	param. Whittle estimator	Example 5 in Dahlhaus and Polonik (2009)
6.	$\phi(u,\lambda) = \left(I_{[0,u_0]}(u)-u_0\right) I_{[0,\lambda_0]}(\lambda)$	testing stationarity	Example 10
7.	$\phi(u,\lambda) = \cos(\lambda k)$	stationary covariance	Remark 6

8.	$\phi(u,\lambda) = \nabla f_{\eta_0}(\lambda)^{-1}$	stat. Whittle estimator	Remark 6
9.	$\phi(u,\lambda) = K_T(\lambda_0 - \lambda)$	stationary spectral density	Remark 6

Examples1–4 and 9 are examples with index functions ϕ_T depending on T. More complex examples are nonparametric maximum likelihood estimation under shape restrictions (Dahlhaus and Polonik, 2006), model selection with a sieve estimator (Van Bellegem and Dahlhaus, 2006) and wavelet estimates (Dahlhaus and Neumann, 2001). Moreover, $F_T(\phi)$ occurs with local polynomial fits (Jentsch, 2006, Kim, 2001) and several statistics suitable for goodness of fit testing. These applications are quite involved.

However, applications are limited to quadratic statistics, that is the empirical spectral measure is usually of no help in dealing with nonlinear models. Furthermore, for linear processes, the empirical process only applies without further modification to the (score function and the Hessian of the) likelihood $\mathcal{L}_T^{GW}(\eta)$ and its local variant $\mathcal{L}_T^{GW}(u,\boldsymbol{\theta})$ and the local Whittle likelihood $\mathcal{L}_T^{W}(u,\boldsymbol{\theta})$. It also applies to the exact likelihood $\mathcal{L}_T^{E}(\eta)$ after proving $\nabla\mathcal{L}_T^{GW}(\eta_0) - \nabla\mathcal{L}_T^{E}(\eta_0) = o_p(T^{-1/2})$ (see also Theorem 9 (i)) and the conditional likelihoods $\mathcal{L}_T^{C}(\eta)$ and $\mathcal{L}_T^{C}(u,\boldsymbol{\theta})$ in the tvAR case (see Remark 9 – in the general case this is not clear yet). For the block Whittle likelihood $\mathcal{L}_T^{BW}(\eta)$, it may also be applied after establishing $\nabla\mathcal{L}_T^{GW}(\eta_0) - \nabla\mathcal{L}_T^{BW}(\eta_0) = o_p(T^{-1/2})$. However, this is also not clear yet.

We first state a central limit theorem for $E_T(\phi)$ with index functions ϕ that do not vary with T. We use the assumption of bounded variation in both components of $\phi(u,\lambda)$. Besides the definition in (61), we need a definition in two dimensions. Let

$$V^2(\phi) = \sup\Bigg\{ \sum_{j,k=1}^{\ell,m} |\phi(u_j,\lambda_k) - \phi(u_{j-1},\lambda_k) - \phi(u_j,\lambda_{k-1}) + \phi(u_{j-1},\lambda_{k-1})| \;:\; 0 \le u_0 < \ldots < u_\ell \le 1;\; -\pi \le \lambda_0 < \ldots < \lambda_m \le \pi;\; \ell,m \in \mathbf{N} \Bigg\}.$$

For simplicity, we set

$$\|\phi\|_{\infty,V} := \sup_u V\big(\phi(u,\cdot)\big), \quad \|\phi\|_{V,\infty} := \sup_\lambda V\big(\phi(\cdot,\lambda)\big),$$
$$\|\phi\|_{V,V} := V^2(\phi) \quad \text{and} \quad \|\phi\|_{\infty,\infty} := \sup_{u,\lambda} |\phi(u,\lambda)|.$$

THEOREM 10. *Suppose Assumption 1 holds and let $\phi_1,\ldots,\phi_d$ be functions with $\|\phi_j\|_{\infty,V}$, $\|\phi_j\|_{V,\infty}$, $\|\phi_j\|_{V,V}$ and $\|\phi_j\|_{\infty,\infty}$ being finite ($j = 1,\ldots,d$). Then*

$$\big(E_T(\phi_j)\big)_{j=1,\ldots,d} \xrightarrow{\mathcal{D}} \big(E(\phi_j)\big)_{j=1,\ldots,d}$$

where $\big(E(\phi_j)\big)_{j=1,\dots,d}$ *is a Gaussian random vector with mean* 0 *and*

$$\begin{aligned}\operatorname{cov}\big(E(\phi_j), E(\phi_k)\big) &= 2\pi \int_0^1 \int_{-\pi}^{\pi} \phi_j(u,\lambda)\,[\phi_k(u,\lambda) + \phi_k(u,-\lambda)]\, f^2(u,\lambda)\, d\lambda\, du \\ &\quad + \kappa_4 \int_0^1 \left(\int_{-\pi}^{\pi} \phi_j(u,\lambda_1) f(u,\lambda_1)\, d\lambda_1 \right) \\ &\quad \times \left(\int_{-\pi}^{\pi} \phi_k(u,\lambda_2) f(u,\lambda_2) d\lambda_2 \right) du. \end{aligned} \tag{99}$$

Proof. See Dahlhaus and Polonik (2009), Theorem 2.5. □

Remark 6 (Stationary processes/model mis-specification by stationary models). The classical central limit theorem for the weighted periodogram in the stationary case can be obtained as a corollary: If $\phi(u,\lambda) = \tilde{\phi}(\lambda)$ is time invariant, then

$$F_T(\phi) = \int_{-\pi}^{\pi} \tilde{\phi}(\lambda) \frac{1}{T} \sum_{t=1}^{T} J_T\left(\frac{t}{T}, \lambda\right) d\lambda = \int_{-\pi}^{\pi} \tilde{\phi}(\lambda) I_T(\lambda)\, d\lambda \tag{100}$$

(see (87)), that is, $F_T(\phi)$ is the classical spectral measure in the stationary case with the following applications:

(i) $\phi(u,\lambda) = \tilde{\phi}(\lambda) = \cos \lambda k$ is the empirical covariance estimator of lag k;
(ii) $\phi(u,\lambda) = \tilde{\phi}(\lambda) = 1/4\pi \nabla f_{\boldsymbol{\theta}}^{-1}(\lambda)$ is the score function of the Whittle likelihood.

Theorem 10 gives the asymptotic distribution for these examples - both in the stationary case and in the misspecified case where the true underlying process is only locally stationary. If $\phi(u,\lambda) = \tilde{\phi}(\lambda)$ is a kernel, we obtain an estimate of the spectral density whose asymptotic distribution is a special case of Theorem 11 below (also in the misspecified case). □

We now state a central limit theorem for $F_T(\phi_T) - F(\phi_T)$ with index functions ϕ_T depending on T. In addition, we extend the hitherto definitions to tapered data

$$X_{t,T}^{(h_T)} := h_T\left(\frac{t}{T}\right) \cdot X_{t,T}$$

where $h_T : (0,1] \to [0,\infty)$ is a data taper (with $h_T(\cdot) = I_{(0,1]}(\cdot)$ being the nontapered case). The main reason for introducing data tapers is to include segment estimates - see the discussion below. As before the empirical spectral measure is defined by

$$F_T(\phi) = F_T^{(h_T)}(\phi) := \frac{1}{T} \sum_{t=1}^{T} \int_{-\pi}^{\pi} \phi\left(\frac{t}{T}, \lambda\right) J_T^{(h_T)}\left(\frac{t}{T}, \lambda\right) d\lambda \tag{101}$$

now with the tapered preperiodogram

$$J_T^{(h_T)}\left(\frac{t}{T},\lambda\right)=\frac{1}{2\pi}\sum_{k:1\le[t+1/2\pm k/2]\le T} X^{(h_T)}_{[t+1/2+k/2],T}\,X^{(h_T)}_{[t+1/2-k/2],T}\exp(-i\lambda k) \tag{102}$$

(we mention that in some cases, a rescaling may be necessary for $J_T^{(h_T)}(u,\lambda)$ to become a pre-estimate of $f(u,\lambda)$ – an obvious example for this is $h_T(u)=(1/2)\,I_{(0,1]}(u)$).

$F(\phi)$ is the theoretical counterpart of $F_T(\phi)$

$$F(\phi)=F^{(h_T)}(\phi):=\int_0^1 h_T^2(u)\int_{-\pi}^{\pi}\phi(u,\lambda)f(u,\lambda)d\lambda\,du. \tag{103}$$

Note that (87) also holds with a data taper, that is

$$\frac{1}{T}\sum_{t=1}^{T}J_T^{(h_T)}\left(\frac{t}{T},\lambda\right)=\frac{H_{2,T}}{T}\,I_T^{(h_T)}(\lambda)$$

with the tapered periodogram

$$I_T^{(h_T)}(\lambda):=\frac{1}{2\pi\,H_{2,T}}\left|\sum_{s=1}^{T}X_s^{(h_T)}\exp(-i\lambda s)\right|^2,\quad\text{where}\quad H_{2,T}:=\sum_{t=1}^{T}h_T\Big(\frac{t}{T}\Big)^2. \tag{104}$$

An important special case is $h_T^{(u_0)}(tT):=k\big(\frac{u_0-t/T}{b_T}\big)$ with bandwidth b_T and k having compact support on $[-\frac{1}{2},\frac{1}{2}]$. If $b_T:=N/T$ then $I_T^{(h_T)}(\lambda)=I_T(u_0,\lambda)$ with $I_T(u_0,\lambda)$ as in (19). If in addition $\phi(u,\lambda)=\psi(\lambda)$, we obtain

$$F_T(\phi)=\int_{-\pi}^{\pi}\psi(\lambda)\left(\frac{1}{T}\sum_{t=1}^{T}J_T^{(h_T)}\left(\frac{t}{T},\lambda\right)\right)d\lambda=\frac{H_{2,T}}{T}\int_{-\pi}^{\pi}\psi(\lambda)I_T^{(h_T)}(\lambda)d\lambda.$$

For example for $\psi(\lambda):=\exp i\lambda k$, this is exactly $H_{2,T}/T\;\hat c_T(u_0,k)$ with the tapered covariance estimate from (8). In this case, $H_{2,T}/T$ is proportional to b_T.

The last example suggests to use $1/H_{2,T}$ instead of $1/T$ in (101) as a norming constant. However, this is not always the right choice (as can be seen from case (ii) in Remark 8).

It turns out that in the above situation, the rate of converge of the empirical spectral measure becomes $\sqrt{T}/\rho_2^{(h_T)}(\phi_T)$, where

$$\rho_2^{(h_T)}(\phi):=\left(\int_0^1 h_T^4(u)\int_{-\pi}^{\pi}\phi(u,\lambda)^2\,d\lambda\,du\right)^{1/2}.$$

Therefore, we can embed this case into the situation treated in the last section by studying the convergence of

$$E_T^{(h_T)}\left(\frac{\phi_T}{\rho_2^{(h_T)}(\phi_T)}\right)=\frac{\sqrt{T}}{\rho_2^{(h_T)}(\phi_T)}\Big(F_T(\phi_T)-F^{(h_T)}(\phi_T)\Big).$$

Furthermore, let

$$c_E^{(h_T)}(\phi_j,\phi_k) \coloneqq 2\pi \int_0^1 h_T^4(u) \int_{-\pi}^{\pi} \phi_j(u,\lambda)\left[\phi_k(u,\lambda)+\phi_k(u,-\lambda)\right] f^2(u,\lambda)d\lambda\, du + \kappa_4 \int_0^1 h_T^4(u)\left(\int_{-\pi}^{\pi}\phi_j(u,\lambda_1)f(u,\lambda_1)d\lambda_1\right)\left(\int_{-\pi}^{\pi}\phi_k(u,\lambda_2)f(u,\lambda_2)d\lambda_2\right)du. \tag{105}$$

THEOREM 11. *Suppose that $X_{t,T}$ is a locally stationary process and suitable regularity conditions hold. If the limit*

$$\Sigma_{j,k} \coloneqq \lim_{T\to\infty} \frac{c_E^{(h_T)}(\phi_{Tj},\phi_{Tk})}{\rho_2^{(h_T)}(\phi_{Tj})\,\rho_2^{(h_T)}(\phi_{Tk})} \tag{106}$$

exists for all $j,k=1,\ldots,d$ then

$$\left(\frac{\sqrt{T}}{\rho_2^{(h_T)}(\phi_{Tj})}\left(F_T(\phi_{Tj})-F^{(h_T)}(\phi_{Tj})\right)\right)_{j=1,\ldots,d} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\Sigma). \tag{107}$$

REMARK 7 (Bias). In addition, we have the bias term

$$\frac{\sqrt{T}}{\rho_2^{(h_T)}(\phi_T)}\left(F^{(h_T)}(\phi_T)-\lim_{T\to\infty}F^{(h_T)}(\phi_T)\right).$$

The magnitude of this bias depends on the smoothness of the time varying spectral density. In this section, we usually require conditions such that this bias is of lower order. This is different in Section 3, where the bias has explicitly been investigated. □

REMARK 8 (Typical applications). A typical application of this result is the case of kernel type local estimators, which can be constructed by using kernels, data tapers or a combination of both:

(i) $\phi_T(u,\lambda)=\frac{1}{b_T}K\left(\frac{u_0-u}{b_T}\right)\psi(\lambda)$ $\quad h_T(\cdot)=I_{(0,1]}(\cdot)$

(ii) $\phi_T(u,\lambda)=\frac{1}{b_T}K\left(\frac{u_0-u}{b_T}\right)\psi(\lambda)$ $\quad h_T(u)=I_{[u_0-b_T/2,u_0+b_T/2]}(u)$

(iii) $\phi_T(u,\lambda)=\psi(\lambda)$ $\quad h_T(\frac{t}{T})=k\left(\frac{u_0-t/T}{b_T}\right)$

where $K(\cdot)$ and $k(\cdot)$ are kernel functions and b_T is the bandwidth. If $K(\cdot)=k(\cdot)^2$, then the resulting estimates all have the same asymptotic properties – see below. Dependent on the function $\psi(\lambda)$, this leads to different applications: If we set $\psi(\lambda)=\cos(\lambda k)$, the estimate (iii) is the estimate $\hat{c}_T(u_0,k)$ from (8) and (i) is "almost" the estimate $\tilde{c}_T(u_0,k)$

from (9) (for k even it is exactly the same, for k odd the difference can be treated with the methods mentioned in Remark 5).

We now show how Theorem 11 leads to the asymptotic distribution for these estimates:

(i) If $K(\cdot)$ and $\psi(\cdot)$ are of bounded variation and $b_T \to 0$, $b_T T \to \infty$, then the regularity conditions of Theorem are fulfilled (see Dahlhaus (2009), Remark 3.4). Furthermore,

$$\rho_2^{(h_T)}(\phi_T) = \rho_2(\phi_T) = \Big(\frac{1}{b_T}\int K^2(x)\,dx \int |\psi(\lambda)|^2\,d\lambda\Big)^{1/2} \approx b_T^{-1/2}. \tag{108}$$

For $f(\cdot,\lambda)$ continuous at u_0, we have

$$c_E^{(h_T)}(\phi_{Tj},\phi_{Tk}) \sim \frac{1}{b_T}\int K^2(x)\,dx \Bigg[2\pi \int_{-\pi}^{\pi} \psi_j(\lambda)\big[\psi_k(\lambda)+\psi_k(-\lambda)\big] \times f^2(u_0,\lambda)\,d\lambda + \kappa_4 \Bigg(\int_{-\pi}^{\pi} \psi_j(\lambda_1) f(u_0,\lambda_1)\,d\lambda_1\Bigg) \times \Bigg(\int_{-\pi}^{\pi} \psi_k(\lambda_2) f(u_0,\lambda_2)\,d\lambda_2\Bigg)\Bigg] =: \frac{1}{b_T}\Gamma_{jk}$$

that is (106) is also fulfilled, and we obtain from Theorem 11

$$\sqrt{b_T T}\Big(F_T(\phi_{Tj}) - F^{(h_T)}(\phi_{Tj})\Big)_{j=1,\dots,d} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\Gamma). \tag{109}$$

(ii) The additional taper $h_T(u) = I_{[u_0-b_T/2,u_0+b_T/2]}(u)$ implies that we use only data from the interval $[u_0 - b_T/2, u_0 + b_T/2]$. We obtain in this case

$$\rho_2^{(h_T)}(\phi_T) = \Bigg(\int_0^1 \frac{1}{b_T^2} K\Big(\frac{u_0-u}{b_T}\Big)^2 du \int_{-\pi}^{\pi} |\psi(\lambda)|^2\,d\lambda\Bigg)^{1/2},$$

i.e., we have the same $\rho_2^{(h_T)}(\phi_T)$ as above. Furthermore, $c_E^{(h_T)}(\phi_T,\phi_T)$ is the same. Thus, we obtain the same asymptotic distribution and the same rate of convergence.

(iii) If $K(\cdot) = k(\cdot)^2$, we obtain in this case

$$\frac{1}{b_T}\rho_2^{(h_T)}(\phi_T) = \Bigg(\int_0^1 \frac{1}{b_T^2} K\Big(\frac{u_0-u}{b_T}\Big)^2 du \int_{-\pi}^{\pi} |\psi(\lambda)|^2\,d\lambda\Bigg)^{1/2}$$

i.e., we obtain again the same expression. Furthermore, $1/b_T^2\, c_E^{(h_T)}(\phi_{Tj},\phi_{Tk})$ is the same as $c_E^{(h_T)}(\phi_{Tj},\phi_{Tk})$ above. Thus, we have again the same asymptotic distribution and the same rate of convergence. □

Example 9 (Curve estimation by local quasi-likelihood estimates). Local Whittle estimates on a segment were defined in (17) and discussed in Example 3 (the bias was heuristically derived in Example 2). We now consider the presumably equivalent local quasi-likelihood estimate defined by

$$\hat{\boldsymbol{\theta}}_T^{GW}(u_0) \coloneqq \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_T^{GW}(u_0, \boldsymbol{\theta}) \tag{110}$$

with

$$\mathcal{L}_T^{GW}(u_0, \boldsymbol{\theta}) \coloneqq \frac{1}{4\pi}\frac{1}{T}\sum_{t=1}^{T} \frac{1}{b_T} K\Big(\frac{u_0 - t/T}{b_T}\Big) \int_{-\pi}^{\pi} \Big\{ \log 4\pi^2 f_{\boldsymbol{\theta}}(\lambda) + \frac{J_T(\frac{t}{T}, \lambda)}{f_{\boldsymbol{\theta}}(\lambda)} \Big\}\, d\lambda. \tag{111}$$

(this is a combination of (26) and (96)). The asymptotic normality of the estimate $\hat{\boldsymbol{\theta}}_T^{GW}(u_0)$ is derived in Dahlhaus (2009), Example 3.6. Key steps in the proof are the fact that both the score function and the Hessian matrix can be written in terms of the empirical spectral process leading to a rather simple proof. For example

$$\sqrt{b_T T}\ \nabla_i \mathcal{L}_T\big(u_0, \boldsymbol{\theta}_0(u_0)\big) = \sqrt{b_T T}\Big(F_T(\phi_{T,u_0,i}) - F(\phi_{T,u_0,i})\Big) + o_p(1) \tag{112}$$

where $\phi_{T,u_0,i}(v,\lambda) \coloneqq \frac{1}{b_T} K(u_0 - v/b_T)\ \frac{1}{4\pi}\nabla_i f_{\boldsymbol{\theta}}^{-1}(\lambda)_{|\boldsymbol{\theta}=\boldsymbol{\theta}_0(u_0)}$. Theorem 11 then immediately gives the asymptotic normality of the score function and after some additional considerations also asymptotic normality of $\hat{\boldsymbol{\theta}}_T^{GW}(u_0)$. For details, see Dahlhaus (2009), Example 3.6.

The above estimate corresponds to case (i) in Remark 8. Case (iii) in Remark 8 leads instead to the tapered Whittle estimate $\hat{\boldsymbol{\theta}}_T^{W}(u_0)$ on the segment, since for $h_T^{(u_0)}(t/T) \coloneqq k(u_0 - t/T/b_T)$, we have $I_T^{(h_T)}(\lambda) = I_T(u_0, \lambda)$ with $I_T(u_0,\lambda)$ as in (19). This estimate has the same asymptotic properties provided $k(\cdot)^2 = K(\cdot)$. It's asymptotic properties can now also be derived by using Theorem 11.

Remark 9 (Related estimates). Many estimates are only approximately of the form discussed above, for example, the sum statistic

$$F_T^{\Sigma}(\phi) \coloneqq \frac{2\pi}{T^2}\sum_{t=1}^{T}\sum_{j=1}^{T} \phi\Big(\frac{t}{T}, \lambda_j\Big)\, J_T^{(h_T)}\Big(\frac{t}{T}, \lambda_j\Big) \tag{113}$$

where $\lambda_j = 2\pi j/T$ – or representations in terms of the Fourier-coefficients. Important examples of related estimates are the spectral density estimate (84), the covariance estimates (9) and (10) and the score function of the local least squares tvAR(p) estimate from Example 1. We mention that the central limit theorem in Theorem 11 also holds for several modified estimators. Details and proofs can be found in Dahlhaus (2009), Section 4.

We now briefly mention the exponential inequality. Since this is a nonasymptotic result, it holds regardless whether ϕ depends on T. Let $\rho_{2,T}(\phi) := \big(1/T \sum_{t=1}^{T} \int_{-\pi}^{\pi} \phi(t/T,\lambda)^2 d\lambda\big)^{1/2}$.

THEOREM 12 (**Exponential inequality**). *Under suitable regularity conditions, we have for all* $\eta > 0$

$$P\Big(\Big| \sqrt{T} \left(F_T(\phi) - \mathbf{E} F_T(\phi) \right) \Big| \geq \eta \Big) \leq c_1 \exp\Big(- c_2 \sqrt{\frac{\eta}{\rho_{2,T}(\phi)}} \Big) \tag{114}$$

with some constants c_1, $c_2 > 0$ *independent of* T.

This result is proved in Dahlhaus and Polonik (2009), Theorem 2.7. There exist several versions of this result – for example in the Gaussian case, it is possible to omit the $\sqrt{\cdot}$ in (114) or to prove a Bernstein-type inequality which is even stronger (cf. Dahlhaus and Polonik, 2006, Theorem 4.1).

Subsequently, a maximal inequality, i.e., an exponential inequality for $\sup_{\phi \in \Phi} |E_T(\phi)|$ has been proved in Dahlhaus and Polonik (2009), Theorem 2.9 under conditions on the metric entropy of the corresponding function class Φ. We refer to that paper for details.

With the maximal inequality, tightness of the empirical spectral process can be proved leading to a functional central limit theorem for the empirical spectral process indexed by a function class (cf. Dahlhaus and Polonik, 2009, Theorem 2.11). Furthermore, a Glivenko Cantelli type result for the empirical spectral process can be obtained (Theorem 2.12).

Other applications of the maximal inequality are for example uniform rates of convergence for different estimates. As an example, we now state a uniform convergence result for the local quasi-likelihood estimate $\hat{\boldsymbol{\theta}}_T^{GW}(u_0)$ from (110).

THEOREM 13. *Let* $X_{t,T}$ *be a locally stationary process with* $\mu(\cdot) \equiv 0$. *Under suitable regularity conditions (in particular under the assumption that* $f_{\boldsymbol{\theta}}(\lambda)$ *is twice differentiable in* $\boldsymbol{\theta}$ *with uniformly Lipschitz continuous derivatives in* λ*), we have for* $b_T T >> (\log T)^6$

$$\sup_{u_0 \in [b_T/2, 1-b_T/2]} \big\| \hat{\boldsymbol{\theta}}_T^{GW}(u_0) - \boldsymbol{\theta}_0(u) \big\|_2 = O_p\Big(\frac{1}{\sqrt{b_T T}} + b_T^2 \Big),$$

that is for $b_T \sim T^{-1/5}$ *we obtain the uniform rate* $O_p\Big(T^{-2/5}\Big)$.

PROOF. The result has been proved in Dahlhaus and Polonik (2009), Theorem 4.1. □

Example 10 (Testing for stationarity). Another application of the maximal inequality is the derivation of a functional central limit for the empirical spectral process. A possible application is a test for stationarity. We briefly present the idea – although we clearly mention that the construction below is finally not successful.

The idea for a test of stationarity is to test whether the time varying spectral density $f(u,\lambda)$ is constant in u. This is for example achieved by the test statistic

$$\sqrt{T}\sup_{u\in[0,1]}\sup_{\lambda\in[0,\pi]}\left|F_T(u,\lambda)-u\,F_T(1,\lambda)\right| \tag{115}$$

where

$$F_T(u,\lambda) := \frac{1}{T}\sum_{t=1}^{[uT]}\int_0^{\lambda} J_T\left(\frac{t}{T},\mu\right)d\mu$$

is an estimate of the integrated time frequency spectral density $F(u,\lambda) := \int_0^u\int_0^{\lambda} f(v,\mu)\,d\mu dv$, and

$$u\,F_T(1,\lambda) = u\int_0^{\lambda} I_T(\mu)d\mu$$

is the corresponding estimate of $F(u,\lambda)$ under the hypothesis of stationarity, where $f(v,\mu)=f(\mu)$. Under the hypothesis of stationarity, we have

$$F(u,\lambda)-u\,F(1,\lambda) = \int_0^1\int_0^{\lambda}\big(I_{[0,u]}(v)-u\big)f(\mu)\,d\mu\,dv = 0$$

and, therefore,

$$\sqrt{T}\Big(F_T(u,\lambda)-u\,F_T(1,\lambda)\Big) = E_T(\phi_{u,\lambda})$$

with $\phi_{u,\lambda}(v,\mu)=\big(I_{[0,u]}(v)-u\big)\,I_{[0,\lambda]}(\mu)$. We now need functional convergence of $E_T(\phi_{u,\lambda})$. Convergence of the finite-dimensional distributions follows from Theorem 10 above. Tightness and, therefore, the functional convergence follows from Theorem 2.11 of Dahlhaus and Polonik (2009). As a consequence, we obtain under the null hypothesis

$$\sqrt{T}\left(F_T(u,\lambda)-u\,F_T(1,\lambda)\right)_{u\in[0,1],\lambda\in[0,\pi]} \xrightarrow{\mathcal{D}} E\big(u,\lambda\big)_{u\in[0,1],\lambda\in[0,\pi]}$$

where $E\big(u,\lambda\big)$ is a Gaussian process. If $\kappa_4=0$ (Gaussian case) and $f(\mu)=c$, it can be shown that this is the Kiefer-Müller process. However, for general f, it is a difficult and unsolved task to calculate or estimate the limit distribution and, in particular, the distribution of the test statistic in (115). This may be done by transformations (like U_p - or T_p - type transforms) and/or by finding an adequate bootstrap method.

We mention that Paparoditis (2009, 2010) has given two different solutions of this testing problem.

7. Additional topics and further references

This section gives an overview over additional topics with further references. We concentrate on work which uses the infill asymptotic approach of local stationarity. Even in this case, it is not possible to give a complete overview.

7.1. Locally stationary wavelet processes

There exists a large number of papers on the use of wavelets for modeling locally stationary processes. The first type of application is to estimate the parameter curves via the use of wavelets. This has been mentioned a few times in the above presentation (cf. (28)).

A breakthrough for the application of wavelets to nonstationary processes was the introduction of "locally stationary wavelet processes" by Nason et al. (2000). This class is somehow the counterpart to the representation (60) for locally stationary processes. It also uses a rescaling argument – thus making all methods for these processes accessible to a meaningful asymptotic theory. Locally stationary wavelet processes are processes with the wavelet representation

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \sum_{j=1}^{\infty}\sum_{k=-\infty}^{\infty} w_{j,k;T}\,\psi_{j,k-t}\,\xi_{j,k} \tag{116}$$

where $\{\xi_{j,k}\}$ are a collection of uncorrelated random variables with mean 0 and variance 1, the $\{\psi_{j,t}\}$ are a set of discrete nondecimated wavelets (compactly supported oscillatory vectors with support proportional to 2^j), and $\{w_{j,k;T}\}$ are a collection of amplitudes that are smooth in a particular way as a function of k. The smoothness of $\{w_{j,k;T}\}$ controls the degree of local stationarity of $X_{t,T}$. The spectrum is linked to the process by $\{w_{j,k;T}\} \approx S_j\left(\frac{k}{T}\right)$. Nason et al. (2000) also define the "evolutionary wavelet spectrum" and show how this can be estimated by a smoothed wavelet periodogram. In addition, this leads to an estimate of the local covariance. An introduction to LSW processes and an overview on early results for such processes can be found in Nason and von Sachs (1999). Fryzlewicz and Nason (2006) suggest the use of a Haar-Fisz method for the estimation of evolutionary wavelet spectra by combining Haar wavelets and the variance stabilizing Fisz transform. Van Bellegem and von Sachs (2008) consider wavelet processes whose spectral density function changes very quickly in time. By using a wavelet-type transform of the autocovariance function with respect to so-called autocorrelation wavelets, they propose a pointwise adaptive estimator of the time varying autocovariance and the time varying spectrum.

Furthermore, several papers mentioned below use the framework of LSW processes.

7.2. Multivariate locally stationary processes

We first mention that, in particular, the Gaussian likelihood theory for locally stationary processes from Section 5 also holds for multivariate processes – see Dahlhaus (2000).

Beyond that Chiann and Morettin (1999, 2005) investigate the estimation of time varying coefficients of a linear system where the input and output are locally stationary

processes. They study different estimation techniques in the frequency and time domain.

Ombao et al. (2001) analyze bivariate nonstationary time series. They use SLEX functions (time-localized generalization of the Fourier waveform) and propose a method that automatically segments the time series into approximately stationary blocks and selects the span to be used to obtain the smoothed estimates of the time varying spectra and coherence. Ombao et al. (2005) use the SLEX framework to build a family of multivariate models that can explicitly characterize the time varying spectral and coherence properties of a multivariate time series. Ombao and Van Bellegem (2008) estimate the time varying coherence by using time-localized linear filtering. Their method automatically selects via tests of homogeneity the optimal window width for estimating local coherence.

Motta et al. (2011) propose a locally stationary factor model for large cross-section and time dimensions. Factor loadings are estimated by the eigenvectors of a nonparametrically estimated covariance matrix. Eichler et al. (2011) investigate dynamic factor modeling of locally stationary processes. They estimate the common components of the dynamic factor model by the eigenvectors of an estimator of the time varying spectral density matrix. This can also be seen as a time varying principal components approach in the frequency domain.

Cardinali and Nason (2010) introduce the concept of costationary of two locally stationary time series where some linear combination of the two processes is stationary. They show that costationarity imply a error-correction type of formula in which changes in the variance of one series are reflected by simultaneous balancing changes in the other. Sanderson et al. (2010) propose a new method of measuring the dependence between nonstationary time series based on a wavelet coherence between two LSW processes.

7.3. Testing of locally stationary processes – In particular tests for stationarity

Among the large literature on testing, there is a considerable part devoted to testing of stationarity. Tests of stationarity have already been proposed and theoretically investigated before the framework of local stationarity was created. In that cases, the theoretical investigations mainly consisted in the investigation of the asymptotic distribution of the test statistics under the hypothesis of stationarity.

Priestley and Subba Rao (1969) proposed testing the homogeneity of a set of evolutionary spectra evaluated at different points of time. For Gaussian processes and for the purpose of a change-point detection, Picard (1985) developed a test based on the difference between spectral distribution functions estimated on different parts of the series and evaluated using a supremum type statistic. Giraitis and Leipus (1992) generalized this approach to the case of linear processes. von Sachs and Neumann (2000) developed a test of stationarity based on empirical wavelet coefficients estimated using localized versions of the periodogram. Paparoditis (2009) developed a nonparametric test for stationarity against the alternative of a smoothly time varying spectral structure based on a local spectral density estimate. He also investigated the power under the fixed alternative of a locally stationary processes. Paparoditis (2010) tested the assumption of stationarity by evaluating the supremum over time of an L_2-distance between the local periodogram over a rolling segment and an estimator of the spectral density obtained

using the entire time series at hand. The critical values of a supremum type test are obtained using a stationary bootstrap procedure. Dwivedi and Subba Rao (2011) construct a Portmanteau type test statistic for testing stationarity of a time series by using the property that the discrete Fourier transforms of a time series at the canonical frequencies are asymptotically uncorrelated if and only if the time series is second-order stationary.

Tests of general hypothesis are derived in Sakiyama and Taniguchi (2003) who test parametric composite hypothesis by the Gaussian likelihood ratio test, the Wald test, and the Lagrange multiplier test. Sergides and Paparoditis (2009) develop tests of the hypothesis that the time varying spectral density has a semiparametric structure. The test introduced is based on a L_2-distance measure in the spectral domain. As a special case, they test for the presence of a tvAR model. A bootstrap procedure is applied to approximate more accurately the distribution of the test statistic under the null hypothesis. Preuß et al. (2011) also test semiparametric hypotheses. Their method is based on an empirical version of the L_2-distance between the true time varying spectral density and its best approximation under the null hypothesis.

Zhou and Wu (2010) construct simultaneous confidence tubes for time varying regression coefficients in functional linear models. Using a Gaussian approximation result for nonstationary multiple time series, they show that the constructed simultaneous confidence tubes have asymptotically correct nominal coverage probabilities.

7.4. Bootstrap methods for locally stationary processes

Bootstrap methods are in particular needed to derive the asymptotic distribution of test statistics. A time domain local block bootstrap procedure for locally stationary processes has been proposed by Paparoditis and Politis (2002) and Dowla et al. (2003). Sergides and Paparoditis (2008) develop a method to bootstrap the local periodogram. Their method generates pseudolocal periodogram ordinates by combining a parametric time and nonparametric frequency domain bootstrap approach. They first fit locally a time varying autoregressive model to capture the essential characteristics of the underlying process. A locally calculated nonparametric correction in the frequency domain is then used so as to improve upon the locally parametric autoregressive fit. Kreiss and Paparoditis (2011) propose a nonparametric bootstrap method by generating pseudo-time series which mimic the local second- and fourth-order moment structures of the underlying process. They prove a bootstrap central limit theorem for a general class of preperiodogram based statistics.

7.5. Model mis-specification and model selection

Model selection criteria have been heuristically suggested many times for time varying processes – cf. Ozaki and Tong (1975), Kitagawa and Akaike (1978), and Dahlhaus (1996b, 1997), among others – in all papers AIC-type criteria have been suggested for different purposes.

Van Bellegem and Dahlhaus (2006) consider semiparametric estimation and estimate the Kullback-Leibler distance between the semiparametric model and the true process. They use this estimate then as a model selection criterion. Hirukawa et al. (2008) propose a generalized information criterion based on nonlinear functionals

of the time varying spectral density. Chandler (2010) investigates how time varying parameters affect order selection.

Another interesting aspect is that many results of this paper also hold under model – misspecification – for example Theorem 8 and the corresponding result for the Block Whittle estimate from (20). An important example is the case where a stationary model is fitted and the underlying process in truth is only locally stationary – see Example 8 and the more detailed discussion for stationary Yule-Walker estimates in Dahlhaus (1997) Section 5.

7.6. *Likelihood theory and large deviations*

LAN is derived in the parametric Gaussian case by Dahlhaus (1996b, 2000) (cf. Remark 3.3 in that paper). A nonparametric LAN result is derived by Sakiyama and Taniguchi (2003) and a LAN result under non-Gaussianity in Hirukawa and Taniguchi (2006). In both papers, the results are applied to asymptotically optimal estimation and testing. For some statistics, also the asymptotic distribution under contiguous alternatives is derived. Tamaki (2009) studies second-order asymptotic efficiency of appropriately modified maximum likelihood estimators for Gaussian locally stationary processes.

Large deviations principles for quadratic forms of locally stationary processes are derived in Zani (2002) including applications to local spectral density and covariance estimation. Wu and Zhou (2011) obtain an invariance principle for nonstationary vector-valued stochastic processes. They show that the partial sums of nonstationary processes can be approximated on a richer probability space by sums of independent Gaussian random vectors.

7.7. *Recursive estimation*

Recursive estimation algorithms are of the form

$$\widehat{\theta}_t = \widehat{\theta}_{t-1} + \lambda_t \ \psi(\boldsymbol{X}_t, \widehat{\theta}_{t-1}) \tag{117}$$

where $\boldsymbol{X}_t = (X_1, \ldots, X_t)'$. The recursive structure yields an update of the estimate as soon as the next observation becomes available, and the estimate, therefore, is particularly of importance in an online situation. For stationary processes, the algorithm is used with $\lambda_t \sim 1/t$, while in nonstationary situations, one uses a nondecreasing λ (constant step-size case) that is the estimate puts stronger weights on the last observations.

Adaptive estimates of the above type have been investigated over the last 30 years in different scientific communities: by system theorists under the name "recursive identification methods" (cf. Ljung, 1977; Ljung and Söderström, 1983), in the stochastic approximation community (cf. Benveniste et al., 1990; Kushner and Yin, 1997), in the neural network community under the name "back-propagation algorithm" (cf. White (1992) or Haykin (1994)), and in applied sciences, particularly for biological and medical applications (cf. Schack and Grieszbach, 1994).

The properties of recursive estimation algorithms have rigorously been investigated in many papers under the premise that the underlying true process is stationary. However, for nonstationary processes and the constant step-size case, there did not

exist for a long time a reasonable framework to study theoretically the properties of these algorithms. This has changed with the concept of locally stationary processes with it's infill asymptotics which now allows for theoretical investigations of these algorithms.

In Moulines et al. (2005), the properties of recursive estimates of tvAR processes have been investigated in the framework of locally stationary processes. The asymptotic properties of the estimator have been proved including a minimax result. In Dahlhaus and Subba Rao (2007), a recursive algorithm for estimating the parameters of a tvARCH process has been proposed. Again, the asymptotic properties of the estimator have been proved.

7.8. Inference for the mean curve

Modeling the time varying mean of a locally stationary process is an important task which has not been discussed in this overview. In principle, nearly all known techniques from nonparametric regression may be used such as kernel estimates, local polynomial fits, wavelet estimates, or others. The situation is, however, much more challenging, since the "residuals" are in this case a locally stationary process which usually is modeled at the same time.

In general, the topic needs more investigation. Dahlhaus (1996a,b, 1997, 2000) and Dahlhaus and Neumann (2001) contain also results where the mean is time varying and/or estimated. A more detailed investigation is contained in Tunyavetchakit (2010) in the context of time varying AR(p) processes, where the mean curve is estimated in parallel and the optimal segment length is determined similar to (16).

7.9. Piecewise constant models

Davis et al. (2005) consider the problem of modeling a class of nonstationary time series using piecewise constant AR processes. The number and locations of the piecewise AR segments, as well as the orders of the respective AR processes, are determined by the minimum description length principle. The best combination is then determined by a genetic algorithm. In Davis et al. (2008) to general parametric time series models for the segments and illustrate the method with piecewise GARCH models, stochastic volatility, and generalized state-space models.

Locally constant parametric models have also been considered in a nonasymptotic approach by Mercurio and Spokoiny (2004) and others where the so-called small modeling bias condition is used to determine the length of the interval of time homogeneity and to fit the parameters – for more details see also Spokoiny (2010).

7.10. Long-memory processes

Beran (2009) and Palma and Olea (2010) have extended the concept of local stationarity to long-range dependent processes. Whereas Beran (2009) uses a nonparametric approach with a local least-squares estimate similar to (26). Palma and Olea (2010) use a parametric approach and use the block Whittle likelihood from (21). Both papers then investigate the asymptotic properties. Roueff and von Sachs (2011) use a local

log-regression wavelet estimator of the time-dependent long-memory parameter and study it's asymptotic properties.

7.11. Locally stationary random fields

Fuentes (2001) studies different methods for locally stationary isotropic random fields with parameters varying across space. In particular, she uses local Whittle estimates. Eckley et al. (2010) propose the modeling and analysis of image texture by using an extension of a locally stationary wavelet process model for lattice processes. They construct estimates of a spatially localized spectrum and a localized autocovariance which are then used to characterize textures in a multiscale and spatially adaptive way. Anderes and Stein (2011) develop a weighted local likelihood estimate for the parameters that govern the local spatial dependency of a locally stationary random field.

7.12. Discrimination analysis

Discrimination analysis for locally stationary processes based on the Kullback-Leibler divergence as a classification criterion has been investigated in Sakiyama and Taniguchi (2004) and for multivariate processes in Hirukawa (2004). Huang et al. (2004) propose a discriminant scheme based on the SLEX-library and a discriminant criterion that is also related to the Kullback-Leibler divergence. Chandler and Polonik (2006) develop methods for the discrimination of locally stationary processes based on the shape of different features. In particular, they use shape measures of the variance function as a criterion for discrimination and apply their method to the discrimination of earthquakes and explosions. Fryzlewicz and Ombao (2009) use a bias-corrected nondecimated wavelet transform for classification in the framework of LSW processes.

7.13. Prediction

Fryzlewicz et al. (2003) address the problem of how to forecast nonstationary time series by means of nondecimated wavelets. Using the class of LSW processes, they introduce a new predictor based on wavelets and derive the prediction equations as a generalization of the Yule-Walker equations. Van Bellegem and von Sachs (2004) apply locally stationary processes to the forecasting of several economic data sets such as returns and exchange rates.

7.14. Finance

There is a growing interest in finance for models with time varying parameters. An overview on locally stationary volatility models is given in Van Bellegem (2011). A general discussion on local stationary in different areas of finance can be found in the study by Guégan (2007) – see also Taniguchi et al. (2008). For example, many researchers are convinced that the observed slow decay of the sample autocorrelation function of absolute stock returns is not a long-memory effect but due to nonstationary changes in the unconditional variance (cf. Fryzlewicz et al., 2006; Mikosch and Stărică, 2004; Stărică and Granger, 2005) leading for example to GARCH models with time varying parameters.

References for work on tvGARCH models have been given in Section 3. Other work on applications of locally stationary processes in finance is, for example, the work on optimal portfolios with locally stationary returns of assets by Shiraishi and Taniguchi (2007). Hirukawa (2006) uses locally stationary processes for a clustering problem of stock returns. Fryzlewicz (2005) models some stylized facts of financial log returns by LSW processes. Fryzlewicz et al. (2006) consider a locally stationary model for financial log-returns and propose a wavelet thresholding algorithm for volatility estimation, in which Haar wavelets are combined with the variance-stabilizing Fisz transform.

7.15. Further topics

Robinson (1989) uses also the infill asymptotics approach in his work on *nonparametric regression* with time varying coefficients. Orbe et al. (2000) estimate nonparametrically a time varying coefficients model allowing for seasonal and smoothness constraints. Orbe et al. (2005) estimate the time varying coefficients under shape restrictions over and for locally stationary regressors. Chiann and Morettin (2005) investigate the estimation of coefficient curves in time varying linear systems.

Estimation of time varying quantile curves for nonstationary processes has been done in Draghicescu et al. (2009) and Zhou and Wu (2009). Specification tests of time varying quantile curves have been investigated in Zhou (2010).

Acknowledgment

The author is grateful to Suhasini Subba Rao for her helpful comments on an earlier version, which lead to significant improvements.

References

Amado, C., Teräsvirta, T., 2011. Modelling volatility with variance decomposition. CREATES Research Paper 2011-1, Aarhus University.

Anderes, E.B., Stein, M.L., 2011. Local likelihood estimation for nonstationary random fields. J. Multivar. Anal. 102, 506–520.

Benveniste, A., Metivier, M., Priouret, P., 1990. Adaptive Algorithms and Stochastic Approximations. Springer-Verlag, Berlin.

Beran, J., 2009. On parameter estimation for locally stationary long-memory processes. J. Stat. Plann. Inference 139, 900–915.

Berkes, I., Horváth, L., Kokoskza, P., 2003. GARCH processes: structure and estimation. Bernoulli 9, 201–207.

Brillinger, D.R., 1981. Time Series: Data Analysis and Theory. Holden Day, San Francisco.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods, 2nd ed. Springer-Verlag, New York.

Cardinali, A., Nason, G., 2010. Costationarity of locally stationary time series. J. Time Ser. Econom. 2, (2), Article 1. doi:10.2202/1941-1928.1074

Chandler, G., 2010. Order selection for heteroscedastic autoregression: A study on concentration. Stat. Prob. Lett. 80, 1904–1910.

Chandler, G., Polonik, W., 2006. Discrimination of locally stationary time series based on the excess mass functional. J. Am. Stat. Assoc. 101, 240–253.

Chiann, C., Morettin, P., 1999. Estimation of time varying linear systems. Stat. Inference Stoch. Proc. 2, 253–285.

Chiann, C., Morettin, P., 2005. Time-domain estimation of time-varying linear systems. J. Nonpar. Stat. 17, 365–383.

Dahlhaus, R., 1988. Empirical spectral processes and their applications to time series analysis. Stoch. Proc. Appl. 30, 69–83.

Dahlhaus, R., 1996a. On the Kullback-Leibler information divergence for locally stationary processes. Stoch. Proc. Appl. 62, 139–168.

Dahlhaus, R., 1996b. Maximum likelihood estimation and model selection for locally stationary processes. J. Nonpar. Stat. 6, 171–191.

Dahlhaus, R., 1996c. Asymptotic statistical inference for nonstationary processes with evolutionary spectra. In: Robinson, P.M., Rosenblatt, M. (Eds.), Athens Conference on Applied Probability and Time Series, Vol II. Lecture Notes in Statistics, Vol. 115, Springer, New York, pp. 145–159.

Dahlhaus, R., 1997. Fitting time series models to nonstationary processes. Ann. Stat. 25, 1–37.

Dahlhaus, R., 2000. A likelihood approximation for locally stationary processes. Ann. Stat. 28, 1762–1794.

Dahlhaus, R., 2009. Local inference for locally stationary time series based on the empirical spectral measure. J. Econom. 151, 101–112.

Dahlhaus, R., Giraitis, L., 1998. On the optimal segment length for parameter estimates for locally stationary time series. J. Time Ser. Anal. 19, 629–655.

Dahlhaus, R., Neumann, M.H., 2001. Locally adaptive fitting of semiparametric models to nonstationary time series. Stoch. Proc. Appl. 91, 277–308.

Dahlhaus, R., Neumann, M.H., von Sachs, R., 1999. Nonlinear wavelet estimation of time-varying autoregressive processes. Bernoulli 5, 873–906.

Dahlhaus, R., Polonik, W., 2006. Nonparametric quasi maximum likelihood estimation for Gaussian locally stationary processes. Ann. Stat. 34, 2790–2824.

Dahlhaus, R., Polonik, W., 2009. Empirical spectral processes for locally stationary time series. Bernoulli 15, 1–39.

Dahlhaus, R., Subba Rao, S., 2006. Statistical inference for locally stationary ARCH models. Ann. Stat. 34, 1075–1114.

Dahlhaus, R., Subba Rao, S., 2007. A recursive online algorithm for the estimation of time-varying ARCH parameters. Bernoulli 13, 389–422.

Davis, R.A., Lee, T., Rodriguez-Yam, G., 2005. Structural break estimation for nonstationary time series models. J. Am. Stat. Assoc. 101, 223–239.

Davis, R.A., Lee, T., Rodriguez-Yam, G., 2008. Break detection for a class of nonlinear time series models. J. Time Ser. Anal. 29, 834–867.

Dowla, A., Paparoditis, E., Politis, D.N., 2003. Locally stationary processes and the local bootstrap. In: Akritas, M.G., Politis, D.N. (Eds.), Recent Advances and Trends in Nonparametric Statistics . Elsevier Science B.V., Amsterdam, 437–445.

Draghicescu, D., Guillas, S., Wu, W.B., 2009. Quantile curve estimation and visualization for non-stationary time series. J. Comput. Graph. Stat. 18, 1–20.

Dunsmuir, W., 1979. A central limit theorem for parameter estimation in stationary vector time series and its application to models for a signal observed with noise. Ann. Stat. 7, 490–506.

Dwivedi, Y., Subba Rao, S., 2011. A test for second-order stationarity of a time series based on the discrete Fourier transform. J. Time Ser. Anal. 32, 68–91.

Dzhaparidze, K., 1971. On methods for obtaining asymptotically efficient spectral parameter estimates for a stationary Gaussian process with rational spectral density. Theory Probab. Appl. 16, 550–554.

Dzhaparidze, K., 1986. Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series. Springer-Verlag, New York.

Eckley, I.A., Nason, G.P., Treloar, R.L., 2010. Locally stationary wavelet fields with application to the modelling and analysis of image texture. Appl. Statist. 59, 595–616.

Eichler, M., Motta, G., von Sachs, R., 2011. Fitting dynamic factor models to non-stationary time series. J. Econom. 163, 51–70.

Fay, G., Soulier, P., 2001 The periodogram of an i.i.d. sequence. Stoch. Proc. Appl. 92, 315–343.

Fox, R., Taqqu, M.S., 1986. Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. Ann. Stat. 14, 517–532.

Fryzlewicz, P., 2005 Modelling and forecasting financial log-returns as locally stationary wavelet processes. J. Appl. Stat. 32, 503–528.

Fryzlewicz, P., Nason, G.P., 2006. Haar-Fisz estimation of evolutionary wavelet spectra. J. R. Stat. Soc. B 68, 611–634.

Fryzlewicz, P., Ombao, H., 2009. Consistent classification of nonstationary time series using stochastic wavelet representations. J. Am. Stat. Assoc. 104, 299–312.

Fryzlewicz, P., Sapatinas, T., Subba Rao, S., 2006. A Haar-Fisz technique for locally stationary volatility estimation. Biometrika 93, 687–704.

Fryzlewicz, P., Sapatinas, T., Subba Rao, S., 2008. Normalised least-squares estimation in time-varying ARCH models. Ann. Stat. 36, 742–786.

Fryzlewicz, P., Subba Rao, S., 2011. On mixing properties of ARCH and time-varying ARCH processes. Bernoulli 17, 320–346.

Fryzlewicz, P., Van Bellegem, S., von Sachs, R., 2003. Forecasting non-stationary time series by wavelet process modeling. Ann. Inst. Stat. Math. 55, 737–764.

Fuentes, M., 2001. A high frequency kriging approach for non-stationary environmental processes. Environmetrics 12, 469–483.

Giraitis, L., Leipus, R., 1992. Testing and estimating in the change-point problem of the spectral function. Lith. Math. J. 32, 15–29.

Granger, C.W.J., Hatanaka, M., 1964. Spectral Analysis of Economic Time Series. Princeton University Press, Princeton.

Grenander, U., Szegö, G., 1958. Toeplitz Forms and their Applications. University of California Press, Berkeley.

Grenier, Y., 1983. Time dependent ARMA modelling of nonstationary signals. IEEE Trans. Acoust. Speech Signal Process. 31, 899–911.

Guégan, D., 2007. Global and local stationary modelling in finance: Theory and empirical evidence. Centre dŠEconomique de la Sorbonne.

Guo, W., Dai, M., Ombao, H.C., von Sachs, R., 2003. Smoothing spline ANOVA for time-dependent spectral analysis. J. Am. Stat. Assoc. 98, 643–652.

Hannan, E.J., 1973. The asymptotic theory of linear time series models. J. Appl. Prob. 10, 130–145.

Hallin, M., 1986. Nonstationary q-dependent processes and time-varying moving average models: invertibility properties and the forecasting problem. Adv. Appl. Probab. 18, 170–210.

Hirukawa, J., 2004. Discriminant analysis for multivariate non-Gaussian locally stationary processes. Sci. Math. Japonicae Online 10, 235–258.

Hirukawa, J., 2006. Cluster analysis for non-Gaussian locally stationary processes. Int. J. Theor. Appl. Fin. 9, 113–132.

Hirukawa, J., Kato, H.S., Tamaki, K., Taniguchi, M., 2008. Generalized information criteria in model selection for locally stationary processes. J. Japan Stat. Soc. 38, 157–171.

Hirukawa, J., Taniguchi, M., 2006. LAN theorem for non-Gaussian locally stationary processes and its applications. J. Stat. Plan. Infer. 136, 640–688.

Hosoya, Y., Taniguchi, M., 1982. A central limit theorem for stationary processes and the parameter estimation of linear processes. Ann. Stat. 10, 132–153.

Huang, H.-Y., Ombao, H.C., Stoffer, D.S., 2004. Discrimination and Classification of Nonstationary Time Series Using the SLEX Model. J. Amer. Stat. Assoc. 99, 763–774.

Jentsch, C., 2006. Asymptotik eines nicht-parametrischen Kernschätzers für zeitvariable autoregressive Prozesse. Diploma thesis, University of Braunschweig.

Kayhan, A., El-Jaroudi, A., Chaparro, L., 1994. Evolutionary periodogram for nonstationary signals. IEEE Trans. Signal Process. 42, 1527–1536.

Kim, W., 2001. Nonparametric kernel estimation of evolutionary autoregressive processes. Discussion paper 103. Sonderforschungsbereich 373, Berlin.

Kitagawa, G., Akaike, H., 1978. A Procedure for The Modeling of Non-Stationary Time Series. Ann. Inst. Stat. Math. 30 B, 351–363.

Kitagawa, G., Gersch, W., 1985. A smoothness priors time-varying AR coefficient modeling of the nonstationary covariance time series. IEEE Trans. Automat. Cntrl. 30, 48–56.

Koo, B., Linton, O., 2010. Semiparametric estimation of locally stationary diffusion models. LSE STICERD Research Paper No. EM/2010/551.

Kreiss, J.-P., Paparoditis, E., 2011. Bootstrapping Locally Stationary Processes. Technical report.

Kushner, H.J., Yin, G.G., 1997. Stochastic Approximation Algorithms and Applications. Springer-Verlag, New York.

Ljung, L., 1977. Analysis of recursive stochastic algorithms. IEEE Trans. Automat. Contr. 22, 551–575.
Ljung, L., Söderström, T., 1983. Theory and Practice of Recursive Identification. MIT Press, Cambridge, MA.
Martin, W., Flandrin, P., 1985. Wigner-Ville spectral analysis of nonstationary processes. IEEE Trans. Acoust. Speech Signal Process. 33, 1461–1470.
Mélard, G., Herteleer-de-Schutter, A., 1989. Contributions to evolutionary spectral theory. J. Time Ser. Anal. 10, 41–63.
Mercurio, D., Spokoiny, V., 2004. Statistical inference for time-inhomogenous volatility models. Ann. Stat. 32, 577–602.
Mikosch, T., Gadrich, T., Klüppelberg, C., Adler, R.J., 1995. Parameter estimation for ARMA models with infinite variance innovations. Ann. Stat. 23, 305–326.
Mikosch, T., Norvaisa, R., 1997. Uniform convergence of the empirical spectral distribution function. Stoch. Proc. Appl. 70, 85–114.
Mikosch, T., Stărică, C., 2004. Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. Rev. Econ. Stat. 86, 378–390.
Motta, G., Hafner, C.M., von Sachs, R., 2011. Locally stationary factor models: Identification and nonparametric estimation. Econom. Theory 27(6), 1279–1319. doi:10.1017/S026646661100005, page 1–41.
Moulines, E., Priouret, P., Roueff, F., 2005. On recursive estimation for locally stationary time varying autoregressive processes. Ann. Stat. 33, 2610–2654.
Nason, G.P., von Sachs, R., 1999. Wavelets in time series analysis. Phil. Trans. R. Soc. Lond. A 357, 2511–2526.
Nason, G.P., von Sachs, R., Kroisandt, G., 2000. Wavelet processes and adaptive estimation of evolutionary wavelet spectra. J. R. Stat. Soc. B 62, 271–292.
Neumann, M.H., von Sachs, R., 1997. Wavelet thresholding in anisotropic function classes and applications to adaptive estimation of evolutionary spectra. Ann. Stat. 25, 38–76.
Ombao, H.C., Raz, J.A., von Sachs, R., Malow, B.A., 2001. Automatic statistical analysis of bivariate nonstationary time series. J. Am. Stat. Assoc. 96, 543–560.
Ombao, H.C., Van Bellegem, S., 2008. Evolutionary coherence of nonstationary signals. IEEE Trans. Signal Process. 56, 2259–2266.
Ombao, H.C., von Sachs, R., Guo, W., 2005. The SLEX analysis of multivariate non-stationary time series. J. Am. Stat. Assoc. 100, 519–531.
Orbe, S., Ferreira, E., Rodriguez-Poo, R.M., 2000. A nonparametric method to estimate time varying coefficients. J. Nonparam. Stat. 12, 779–806.
Orbe, S., Ferreira, E., Rodriguez-Poo, R.M., 2005. Nonparametric estimation of time varying parameters under shape restrictions. J. Econom. 126, 53–77.
Ozaki, T., Tong, H., 1975. On the fitting of non-stationary autoregressive models in time series analysis. Proceedings of the 8th Hawaii International Conference on System Sciences. Western Periodical Company, California.
Palma, W., Olea, R., 2010. An efficient estimator for locally stationary Gaussian long-memory processes. Ann. Stat. 38, 2958–2997.
Paparoditis, E., 2009. Testing temporal constancy of the spectral structure of a time series. Bernoulli 15, 1190–1221.
Paparoditis, E., 2010. Validating stationarity assumptions in time series analysis by rolling local periodograms. J. Amer. Statist. Assoc. 105, 839–851.
Paparoditis, E., Politis, D.N., 2002. Local block bootstrap. C. R. Acad. Sci. Paris, Ser. I 335, 959–962.
Parzen, E., 1983. Autoregressive spectral estimation. In: Brillinger, D.R., Krishnaiah, P.R. (Eds.), Handbook of Statistics. North-Holland, Amsterdam, 3, pp. 221–247.
Picard, D., 1985. Testing and estimating change-points in time series. Adv. Appl. Probab. 17, 841–867.
Preuß, P., Vetter, M., Dette, H., 2011. Testing semiparametric hypotheses in locally stationary processes. Discussion paper 13/11. SFB 823, TU Dortmund.
Priestley, M.B., 1965. Evolutionary spectra and non-stationary processes. J. R. Stat. Soc. Ser. B 27, 204–237.
Priestley, M.B., 1981. Spectral Analysis and Time Series. Academic Press, London.
Priestley, M.B., 1988. Nonlinear and Nonstationary Time Series Analysis, Academic Press, London.
Priestley, M.B., Subba Rao, T., 1969. A test for non-stationarity of time series. J. R. Stat. Soc. B 31, 140–149.
Robinson, P.M., 1989. Nonparametric estimation of time varying parameters. In: Hackl, P. (Ed.), Statistics Analysis and Forecasting of Economic Structural Change. Springer, Berlin, pp. 253–264.

Robinson, P.M., 1995. Gaussian semiparametric estimation of long range dependence. Ann. Stat. 23, 1630–1661.

Rosen, O., Stoffer, D.S., Wood, S., 2009. Local Spectral Analysis via a Bayesian Mixture of Smoothing Splines. J. Am. Stat. Assoc. 104, 249–262.

Roueff, F., von Sachs, R., 2011. Locally stationary long memory estimation. Stoch. Proc. Appl. 121, 813–844.

Sanderson, J., Fryzlewicz, P., Jones, M., 2010. Estimating linear dependence between nonstationary time series using the locally stationary wavelet model. Biometrika 97, 435–446.

Sakiyama, K., Taniguchi, M., 2003. Testing composite hypotheses for locally stationary processes. J. Time Ser. Anal. 24, 483–504.

Sakiyama, K., Taniguchi, M., 2004. Discriminant analysis for locally stationary processes. J. Multiv. Anal. 90, 282–300.

Schack, B., Grieszbach, G., 1994. Adaptive methods of trend detection and their application in analyzing biosignals. Biom. J. 36, 429–452.

Sergides, M., Paparoditis, E., 2008. Bootstrapping the Local Periodogram of Locally Stationary Processes. J. Time Ser. Anal. 29, 264–299. Corrigendum: J. Time Ser. Anal. 30, 260–261.

Sergides, M., Paparoditis, E., 2009. Frequency domain tests of semiparametric hypotheses for locally stationary processes. Scandin. J. Stat. 36, 800–821.

Shiraishi, H., Taniguchi, M., 2007. Statistical estimation of optimal portfolios for locally stationary returns of assets. Int. J. Theor. Appl. Finance 10, 129–154.

Spokoiny, V., 2010. Local parametric methods in nonparametric estimation. Springer-Verlag, Berlin Heidelberg New York.

Stărică, C., Granger, C., 2005. Nonstationarities in stock returns. Rev. Econ. Stat. 87, 503–522.

Subba Rao, S., 2006. On some nonstationary, nonlinear random processes and their stationary approximations. Adv. Appl. Probab. 38, 1155–1172.

Subba Rao, T., 1970. The fitting of non-stationary time series models with time-dependent parameters. J. R. Stat. Soc. B 32, 312–322.

Tamaki, K., 2009. Second order properties of locally stationary processes. J. Time Ser. Anal. 30, 145–166.

Taniguchi, M., Kakizawa, Y., 2000. Asymptotic Theory of Statistical Inference for Time Series. Springer Verlag, New York.

Taniguchi, M., Hirukawa, J., Tamaki, K., 2008. Optimal Statistical Inference in Financial Engineering. Chapman & Hall, Boca Raton.

Tjøstheim, D., 1976. Spectral generating operators for non-stationary processes. Adv. Appl. Probab. 8, 831–846.

Tunyavetchakit, S., 2010. On the optimal segment length for tapered Yule-Walker estimates for time-varying autoregressive processes. Diploma Thesis, Heidelberg.

Van Bellegem, S., Dahlhaus, R., 2006. Semiparametric estimation by model selection for locally stationary processes. J. R. Stat. Soc. B 68, 721–764.

Van Bellegem, S., von Sachs, R., 2004. Forecasting economic time series with unconditional time varying variance. Int. J. Forecast. 20, 611–627.

Van Bellegem, S., von Sachs, R., 2008. Locally adaptive estimation of evolutionary wavelet spectra. Ann. Stat. 36, 1879–1924.

Van Bellegem, S., 2011. Locally stationary volatility models. In: Bauwens, L., Hafner, C., Laurent, S. (eds.), Wiley Handbook in Financial Engineering and Econometrics: Volatility Models and Their Applications, Wiley, New York.

Vogt, M., 2011. Nonparametric regression for locally stationary time series. Preprint, University of Mannheim.

von Sachs, R., Neumann, M., 2000. A wavelet-based test for stationarity. J. Time Ser. Anal. 21, 597–613.

White, H., 1992. Artificial Neural Networks. Blackwell, Oxford.

Whittle, P., 1953. Estimation and information in stationary time series. Ark. Mat. 2, 423–434.

Whittle, P., 1954. Some recent contributions to the theory of stationary processes. Appendix to A study in the analysis of stationary time series, by H. Wold, 2nd ed. 196–228. Almqvist and Wiksell, Uppsala.

Wu, W.B., Zhou, Z., 2011. Gaussian approximations for non-stationary multiple time series. Statistica Sinica 21, 1397–1413.

Zani, M., 2002. Large deviations for quadratic forms of locally stationary processes. J. Multivar. Anal. 81, 205–228.

Zhou, Z., 2010. Nonparametric inference of quantile curves for nonstationary time series. Ann. Stat. 38, 2187–2217.

Zhou, Z., Wu, W.B., 2009. Local linear quantile estimation for non-stationary time series. Ann. Stat. 37, 2696–2729.

Zhou, Z., Wu, W.B., 2010. Simultaneous inference of linear models with time varying coefficients. J. R. Stat. Soc. B 72, 513–531.

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00014-4

Analysis of Multivariate Nonstationary Time Series Using the Localized Fourier Library

Hernando Ombao

Department of Statistics, University of California, Irvine, CA 92697, USA

Abstract

Using the SLEX library as the primary tool, we develop a systematic, flexible, and computationally efficient procedure for analyzing multivariate nonstationary time series. The SLEX library is a collection of bases; each of which consists of localized Fourier waveforms. In the problem of signal representation and spectral estimation, one can select, from the set of bases in the SLEX library, the one that best represents the underlying process. Moreover, in discrimination and classification of nonstationary time series, one can select the basis that gives the maximal separation between classes of nonstationary time series. We illustrate the SLEX methods by analyzing multichannel EEGs recorded during an epileptic seizure and during a visual-motor experiment.

Keywords: Coherence, discrimination, fourier transform, nonstationary time series, smooth localized complex exponentials, spectral analysis, spectral matrix.

1. Introduction

Many brain science experiments collect multivariate time series from animal and human subjects to study brain electrical, magnetic, and hemodynamic activity. In this chapter, we present methods for analyzing multivariate time series based on the local Fourier library. One example of multivariate time series is electroencephalograms (EEGs) that are measures of brain electrical activity recorded from many sensors on a scalp. Here, we shall analyze the multichannel EEGs recorded during an epileptic seizure and during a visual-motor task experiment in which the participant moved the joystick either to the left or to the right in response to the stimulus presented.

1.1. Goals in the chapter

In the first data set (see Fig. 1), our goal is to study how the composition of waveforms evolve during an episode of an epileptic seizure. Here, we present a method for estimating the spectra (which measure variance decomposition) and coherence (which measures dynamic cross-relationships in the multivariate time series). The second data set (see Fig. 2) was recorded in an experiment in which participants performed a simple voluntary movement that required quick displacements of a hand-held joystick from a central position either to the right or to the left. It is likely that the nature of interactions between brain regions differs between the "right" and "left" conditions. Here, we present a method for selecting spectral features that discriminate between presumed brain connectivity occurring during leftward and rightward movements, aiming to predict intentions to move by assessing the information evident in an electroencephalogram (EEG) time series recorded contemporary with the voluntary movements.

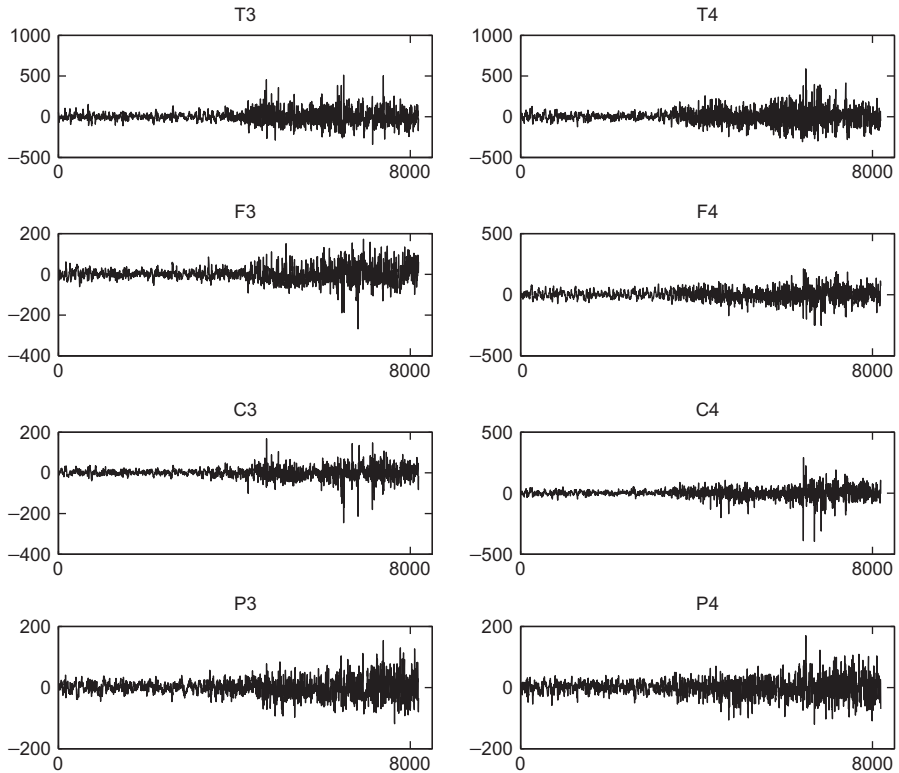

Fig. 1. Electroencephalogram recorded during an epileptic seizure. $T = 8192$. Sampling rate is 100 Hz. Only eight EEG plots are shown although the analysis was conducted on the data set that consists of $p = 18$ channels. The EEG plots on the *left column* are recordings from the *left side of the brain*: T3 (left temporal lobe); F3 (left frontal); C3 (left central); and P3 (left parietal). The plots on the *right column* are recordings from the *right side of the brain*: T4 (right temporal lobe); F4 (right frontal); C4 (right central); and P4 (right parietal).

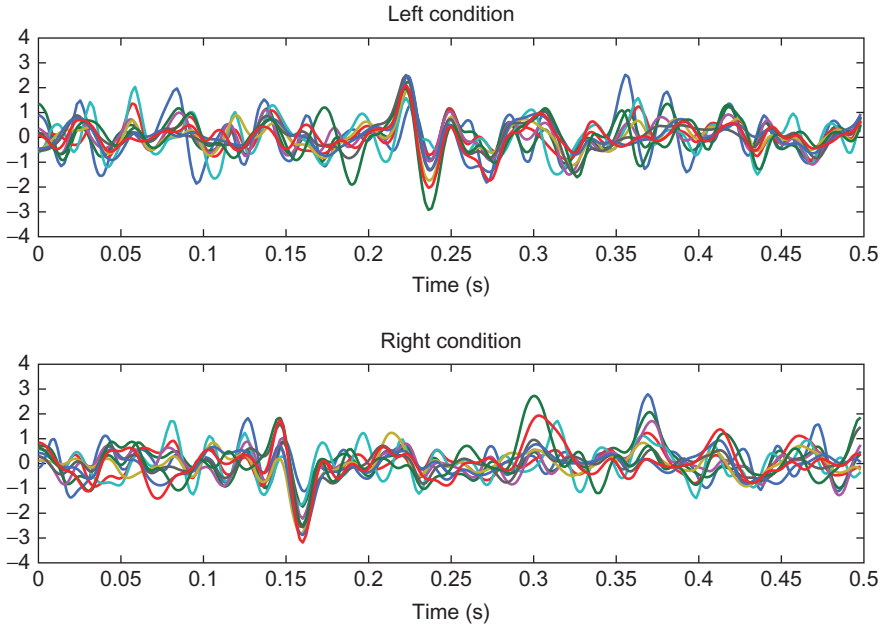

Fig. 2. Left: representative 11-channel EEG recorded from one trial for the *left* condition. Right: representative 11-channel EEG recorded from one trial for the *right* condition.

Most brain time series data are typically nonstationary – their statistical properties and spectral content evolve over time. Such signals cannot be adequately studied using classical Fourier analysis. In this chapter, we shall analyze these brain signals using a **library** of localized Fourier waveforms. This library consists of several bases: for the first goal, we shall estimate the evolutionary spectra and coherence by selecting the basis that gives the best representation for the nonstationary time series, and for the second goal, we select the basis that gives the maximal discrimination or separation between the two classes of signals (leftward vs. rightward movements).

1.2. Overview of spectral representations

We first give a brief background on the spectral representations of time series. Let $\mathbf{X}(t) = [X_1(t), \ldots, X_P(t)]'$ be a zero-mean P-variate time series. Under stationarity, $\mathbf{X}(t)$ can be expressed as a randomly weighted sine and cosine waveform via the Cramér (spectral) representation

$$\mathbf{X}(t) = \int_{-1/2}^{1/2} \mathbf{A}(\omega)\, \exp(\, i2\pi\omega t)\mathbf{dZ}(\omega), \tag{1}$$

where $\mathbf{A}(\omega)$ is the transfer function matrix of dimension $P \times P$ and $d\mathbf{Z}(\omega)$ is a zero-mean orthonormal increment random process, i.e., $\mathbb{C}\mathrm{ov}[d\mathbf{Z}(\omega), d\mathbf{Z}(\lambda)] = \mathbf{0}$ when

$\omega \neq \lambda$ and $\text{Var}[d\mathbf{Z}(\omega)] = \mathbf{1}d\omega$ where $\mathbf{1}$ is the $P \times P$ identity matrix. The spectral density matrix of $\mathbf{X}(t)$ is $\mathbf{f}(\omega) = \mathbf{A}(\omega)\mathbf{A}^*(\omega)$, which is complex-valued Hermitian with dimension $P \times P$.

The relevant spectral quantities are as follows: the autospectrum at frequency ω for $X_p(t)$ is the pth element of the diagonal denoted $f_{pp}(\omega)$; the cross-spectrum between the pth and qth components is the (p, q) element of the spectral matrix denoted $f_{pq}(\omega)$; and the coherence between the pth and qth components is defined to be $\rho_{pq}(\omega) = |f_{pq}(\omega)|^2 / \left[f_{pp}(\omega) f_{qq}(\omega)\right]$. Coherence is a measure of linear association between the ω-oscillatory activity at the p and q components. In fact, Ombao and Van Bellegem (2008) demonstrate that when one applies a linear filter on $X_p(t)$ and $X_q(t)$, so that the resulting filtered signals have spectra concentrated on ω, coherence is approximately equal to the squared cross-correlation between these filtered signals.

Coherence cannot differentiate between direct versus indirect linear association. For example, $X_p(t)$ and $X_q(t)$ may be strongly coherent at some frequency ω but that association may be due to another component $X_r(t)$. Thus, to measure the direct linear associations, one uses partial coherence that is characterized as follows. Define $\mathbf{G}(\omega) = \mathbf{f}^{-1}(\omega)$ and denote its diagonal elements to be $g_{pp}(\omega)$ and define the matrix $\mathbf{H}(\omega)$ to be a diagonal $P \times P$ matrix whose elements are $h_{pp}(\omega) = 1/\sqrt{g_{pp}(\omega)}$. Further, define the matrix $\Lambda(\omega) = -\mathbf{H}(\omega)\mathbf{f}^{-1}(\omega)\mathbf{H}(\omega)$. Partial coherency between the pth and qth components is then defined to be the (p, q)th element of the matrix $\Lambda(\omega)$, denoted by $\Lambda_{pq}(\omega)$, and partial coherence is the square modulus $|\Lambda_{pq}(\omega)|^2$.

Note that for stationary processes: (a) the representation uses the Fourier complex exponentials as the building blocks; (b) the transfer function depends only on frequency and does not vary across time, and consequently, (c) the spectral matrix and all spectral quantities remain constant in time. For nonstationary multivariate time series, Dahlhaus (2001) developed a representation that also uses the Fourier waveforms as building blocks but allows the transfer function to change over time. Here, to present ideas, we use its approximate representation

$$\mathbf{X}_{t,T} \approx \int_{-1/2}^{1/2} \mathbf{A}(t/T, \omega)\ \exp(\, i2\pi\omega t)\mathbf{dZ}(\omega), \tag{2}$$

where $\mathbf{A}(t/T, \omega)$ is the transfer function matrix that is defined on rescaled time t/T. Under this representation, the ingredients are still the Fourier waveforms but the random coefficients $\mathbf{A}(t/T, \omega)\mathbf{dZ}(\omega)$ depend both on time and on frequency. The spectral matrix, defined on rescaled time $u \in [0, 1]$ and frequency $\omega \in (-0.5, 0.5)$, is defined to be $\mathbf{f}(u, \omega) = \mathbf{A}(u, \omega)\mathbf{A}^*(u, \omega)$. The Dahlhaus model provides a asymptotic framework under which one can establish consistency of the estimator for the time-varying spectral matrix.

In this chapter, we shall introduce a complementary approach that utilizes the library of *localized* Fourier waveforms. The remainder of this chapter is organized as follows. The basic ideas on the SLEX waveforms and transform are given in Section 2. The method for fitting the SLEX model and estimating the time-varying spectral properties is given in Sections 3 and 4; we present the procedure for discriminating between classes of nonstationary time series by finding the best SLEX basis that gives the largest separation between the classes.

2. Overview of SLEX analysis

2.1. The SLEX waveforms

The Fourier waveforms are not ideal for analyzing nonstationary time series data because they cannot directly capture the time-localized spectral features of the signal. One standard approach to study nonstationary time series uses windowed Fourier waveforms (Daubechies, 1992) $\phi_F(u) = \Psi(u)\exp(i2\pi\omega u)$, where Ψ is a taper with compact support and $\omega \in (-1/2, 1/2]$. Windowed Fourier waveforms are localized in time but are generally nonorthogonal due to the Balian–Low theorem that states that no smooth window exists so that the windowed Fourier basis functions are simultaneously orthogonal and localized (Wickerhauser, 1994). Orthogonality is a desirable property because it provides elegant representations and gives a unique time–frequency decomposition. Orthogonal transforms preserve the energy of the time series and allow the use of the best basis algorithm (BBA) of Coifman and Wickerhauser (1992) which is computationally efficient and hence facilitates the analysis of massive data sets.

While there exist many localized and orthonormal basis functions that could be used for analyzing nonstationary time series (e.g., wavelets and wavelet packets), there is a strong rationale for using time-localized generalizations of the *Fourier* waveforms. Here, following Ombao et al. (2001), we analyze nonstationary time series data using the SLEX (**S**mooth **L**ocalized Complex **EX**ponential) waveforms

$$\phi_\omega(u) = \Psi_+(u)\exp(i2\pi\omega u) + \Psi_-(u)\exp(-i2\pi\omega u), \tag{3}$$

where $\omega \in (-1/2, 1/2]$, and $u \in \mathcal{I} = [-\eta, 1+\eta]$, $0 < \eta < 0.5$. The windows, plotted in Fig. 3, come in pairs. That is, once Ψ_+ is specified, Ψ_- is determined. Moreover, these windows can be compressed or dilated so that the rescaled support $B \subset \mathcal{I}$. Plots of the SLEX waveforms are given in Fig. 4. The SLEX waveforms are simultaneously orthogonal and smooth. Unlike the smooth-windowed Fourier complex exponentials, the SLEX waveforms evade the Balian–Low obstruction because they are constructed using a projection operator rather than a single window. Details on the construction of waveforms from projection operators are provided in the study by Auscher et al. (1992). A comparison of the SLEX and the other transforms is briefly discussed in Section 2.6. The SLEX library is appealing for modeling time series because it naturally captures the time-lag structure between components of a multivariate time series via the phases of the time-varying cross-spectra and gives results that are easy to interpret because they are time-dependent generalizations of Fourier analysis of stationary time series.

2.2. The SLEX library

The SLEX library is a collection of bases; each basis consists of the SLEX waveforms that are localized; thus, they are able to capture the local spectral features of the time series. Moreover, the SLEX library allows a flexible and rich representation of the observed time series. To illustrate these ideas, we construct a SLEX library in Fig. 5 with level $J = 2$. There are seven dyadic time blocks in this library. They are

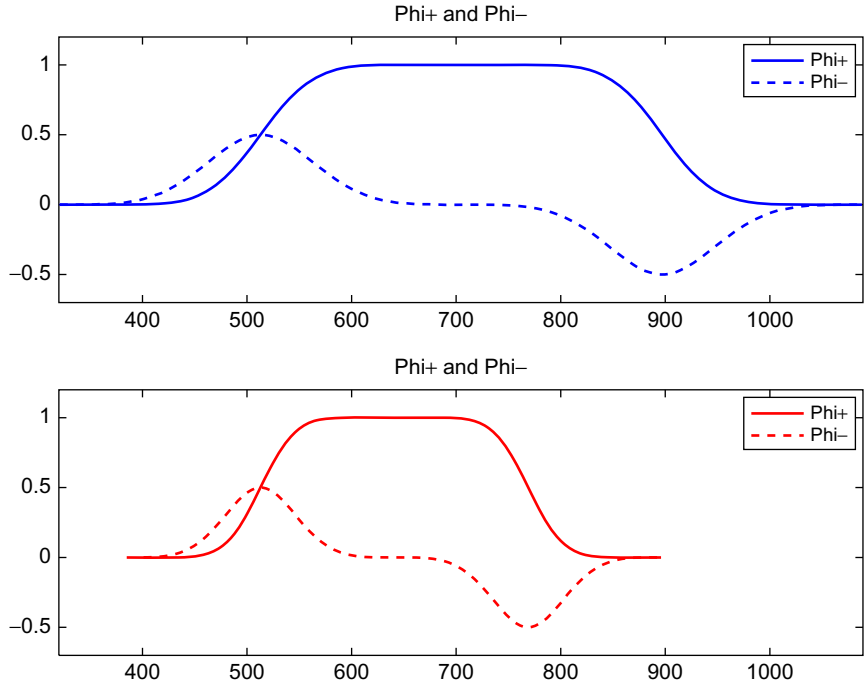

Fig. 3. Smooth window pairs $\Psi_{+,B}(u)$ and $\Psi_{-,\omega,B}(u)$. These windows can be stretched or compressed. In the top picture, B is approximately the rescaled interval $(500/1000, 900/1000)$; in the bottom picture, B is approximately $(500/1000, 750/1000)$.

$S(0,0)$ that covers the entire time series, $S(1,0)$ and $S(1,1)$, which are the two half blocks, and $S(2,b), b = 0,1,2,3$, which are the four quarter blocks. Note that in general, for each resolution level $j = 0,1,\ldots,J$, there are 2^j time blocks, each having length $T/2^j$. We will adopt the notation $S(j,b)$ to denote the block b on level j where $b = 0,1,\ldots,2^j - 1$. The time blocks $S(j,b)$ correspond to the rescaled blocks $B(j,b)$ in the following manner: $S(0,0)$ corresponds to $[0,1]$; $S(1,0)$ corresponds to $(0,1/2)$; $S(2,3)$ corresponds to $[3/4,1]$.

There are five possible bases from this particular SLEX library. One particular basis is composed of blocks $S(1,0), S(2,2), S(2,3)$ that correspond to the shaded blocks in Fig. 5. We point out that each basis is allowed to have multiresolution scales, i.e., a basis can have time blocks with different lengths. This is ideal for processes whose regimes of stationarity have lengths that also vary with time. In choosing the finest time scale (or deepest level) of the transform J, the statistician will need some advice from collaborators who can give some guidance regarding an appropriate time resolution of EEGs. In general, the blocks should be small enough so that we can be confident that the time series is stationary in these blocks. At the same time, the blocks should not be smaller than what is necessary in order to control the variance of the spectral estimator.

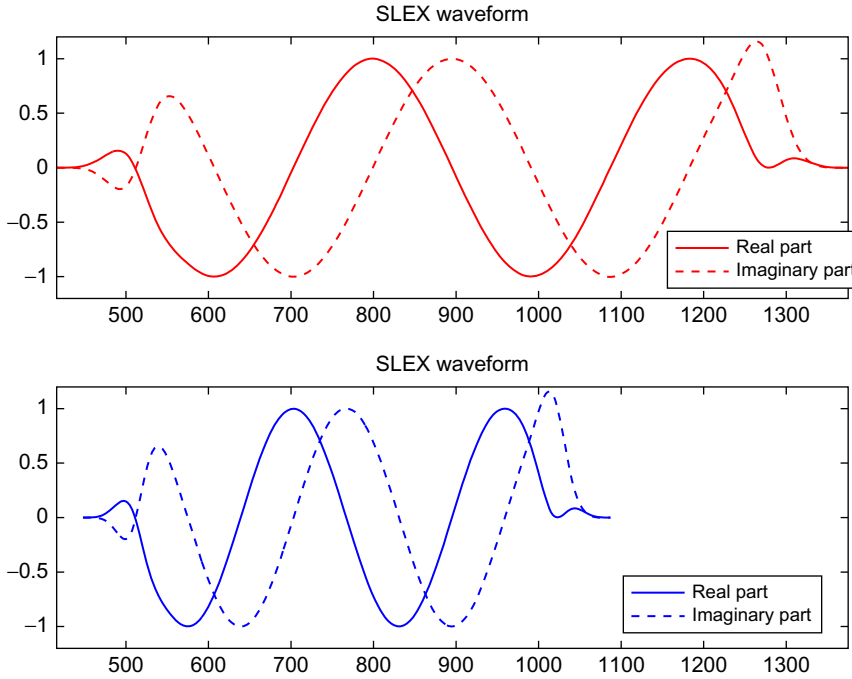

Fig. 4. Examples of the SLEX waveforms at different scales and locations. The SLEX waveforms can be dilated or compressed and then shifted.

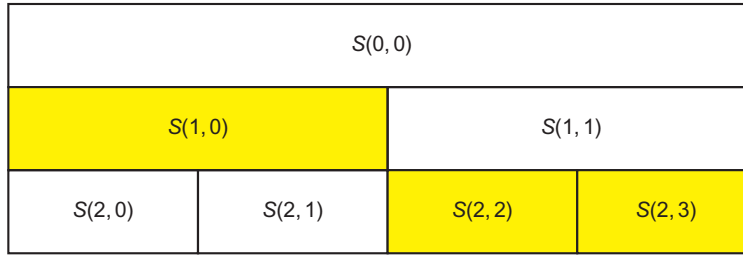

Fig. 5. A SLEX library with level $J = 2$. The shaded blocks represent one basis from the SLEX library.

2.3. Computing the SLEX transform

The SLEX transform is a collection of coefficients corresponding to the SLEX waveforms in the SLEX library. First, we define the SLEX *vector* defined on the discretized time block $S(j, b)$ consisting of time points $\{\alpha_0, \ldots, \alpha_1 - 1\}$. Furthermore, define $|S| = \alpha_1 - \alpha_0$ and the overlap $\epsilon = [\eta\ |S|]$, where [.] denotes the greatest integer function. Since the SLEX waveforms provide smooth transitions across blocks, the SLEX vector is actually defined on an "expanded" block $\{\alpha_0 - \epsilon, \ldots, \alpha_1 - 1 + \epsilon\}$. The SLEX

vector on S oscillating at frequency ω_k takes the form

$$\phi_{S,\omega_k}(t) = \Psi_+\left(\frac{t-\alpha_0}{|S|}\right)\exp(i2\pi\omega_k(t-\alpha_0)) + \Psi_-\left(\frac{t-\alpha_0}{|S|}\right)\exp(-i2\pi\omega_k(t-\alpha_0)),$$

where $\omega_k = k/|S|, k = -\frac{|S|}{2}-1,\ldots,\frac{|S|}{2}$.

Next, we show that the SLEX coefficients can be computed using the fast Fourier transform (FFT). Let $X_\ell(t)$ be one component of a P-channel time series $\mathbf{X}(t)$ of length T. The SLEX coefficients (corresponding to $X_\ell(t)$) on block $S(j,b)$ are defined as:

$$\begin{aligned} d^\ell_{j,b}(\omega_k) &= (M_j)^{-1/2}\sum_t X_\ell(t)\overline{\phi_{j,b,\omega_k}(t)} \\ &= (M_j)^{-1/2}\sum_t \Psi_+\left(\frac{t-\alpha_0}{|S|}\right)X_\ell(t)\,\exp[-i2\pi\omega_k(t-\alpha_0)] \\ &\quad + (M_j)^{-1/2}\sum_t \Psi_-\left(\frac{t-\alpha_0}{|S|}\right)X_\ell(t)\,\exp[i2\pi\omega_k(t-\alpha_0)], \end{aligned}$$

where $M_j = |S(j,b)| = T/2^j$. In the implementation, the "edge" blocks in each level j, namely $S(j,0)$ and $S(j,2^j-1)$, are padded with zeros when we compute the SLEX transform. Finally, by using the FFT, the number of operations needed to compute the SLEX transform has order of magnitude $O[T(\log_2 T)^2]$.

2.4. *Computing the SLEX periodogram matrix*

Denote $\mathbf{d}_{j,b}(\omega_k)$ to be a $P\times 1$ vector of SLEX coefficients at block $S(j,b)$ and frequency ω_k, $\mathbf{d}_{j,b}(\omega_k) = [d^1_{j,b}(\omega_k),\ldots,d^P_{j,b}(\omega_k)]'$. The SLEX periodogram matrix is $\mathbf{I}_{j,b}(\omega_k) = \mathbf{d}_{j,b}(\omega_k)\mathbf{d}^*_{j,b}(\omega_k)$, where $\mathbf{d}^*$ is the complex conjugate transpose of $\mathbf{d}$. The diagonal elements of $\mathbf{I}_{j,b}$ are the SLEX autoperiodograms $I^{\ell\ell}_{j,b}(\omega_k) = |d^\ell_{j,b}(\omega_k)|^2$, while off-diagonal elements are the SLEX cross-periodograms $I^{pq}_{j,b}(\omega_k) = d^p_{j,b}(\omega_k)d^{q*}_{j,b}(\omega_k)$. Analogous to the Fourier periodogram matrices, we also smooth the SLEX periodogram matrices across frequency $\widetilde{\mathbf{I}}_{j,b}(\omega_k) = \frac{1}{2L+1}\sum_{r=-L}^{L}\mathbf{I}_{j,b}(\omega_{k+r})$ to produce a mean-squared consistent estimator.

2.5. *Best basis algorithm*

As already noted, the SLEX library forms a collection of SLEX bases. Depending on the problem at hand, we would like to choose an optimal basis for the purpose of signal representation and another optimal basis for the purpose of signal discrimination. In both cases, we compute some criterion at each time block $S(j,b)$ or rescaled block $B(j,b)$ and denote the value of the criterion to be $\mathcal{C}(j,b)$. Here, we will describe the best basis algorithm (BBA) developed in the study by Coifman and Wickerhauser (1992).

Algorithm

Set maximal level J. Note that the blocks at the finest level will have length $T/2^J$.

Mark the terminal blocks $S(J,0),\ldots,S(J,2^J-1)$.

For $j = J-1,\ldots,0$

For $b = 0,1,\ldots,2^j-1$

Compare the cost at the mother block $S(j,b)$ with that at the children blocks $S(j+1,2b)$ and $S(j+1,2b+1)$.

If $\mathcal{C}(j,b) < \mathcal{C}(j+1,2b) + \mathcal{C}(j+1,2b+1)$ then

Mark the block $S(j,b)$.

Else

Mark the blocks $B(j+1,2b)$ and $B(j+1,2b+1)$ and

Replace $\mathcal{C}(j,b)$ with $\mathcal{C}(j+1,2b) + \mathcal{C}(j+1,2b+1)$.

End b

End j

Finally, the best basis consists of the marked blocks whose ancestors are unmarked.

2.6. The SLEX and other localized waveforms

Wavelets are mathematical functions with localized oscillatory features. They are commonly utilized to estimate functions that have sudden bursts and peaks at localized regions. See the study by Donoho and Johnstone (1994, 1995) for seminal work on nonparametric function estimation using wavelets. Moreover, wavelets have also been developed in Nason et al. (2000) for representing time series with time-varying spectral and scale decompositions. For a comprehensive treatment on the applications of wavelets to various statistical problems, see the study by Vidakovic (1999).

Wavelet packets form another class of localized waveforms. Wavelet packets are a generalization of wavelets. Analogously to SLEX, wavelet packets also form a library of orthonormal bases, which contains the wavelet basis. Due to their generality, wavelet packets offer more flexibility in representing signals that exhibit oscillatory or periodic behavior. One distinction between wavelet packets and SLEX is the manner in which they segment the time–frequency plane. Although the SLEX library is obtained by generating waveforms whose time support dyadically divides the time axis, the wavelet packet library consists of waveforms whose spectral power or frequency support dyadically divides the frequency axis. The best wavelet packet orthonormal basis can also be selected using the best basis algorithm. Detailed discussion on the construction of the wavelet packets is given in the study by Wickerhauser (1994).

Cosine packets are also another time and frequency localized waveforms (Auscher et al., 1992). The cosine packet transform (CPT) shares common features with the SLEX transform: both are time-localized trigonometric functions and both dyadically divide the time axis (rather than the frequency axis, as in WPT) of the time–frequency plane. Moreover, both waveforms are obtained by applying the same window pairs. An application of Ψ_+ and Ψ_- (see Fig. 3) on the complex exponential function produces the SLEX waveform, whereas an application of the same windows on the cosine functions gives the cosine packet waveforms.

There are a number of advantages to using the SLEX rather than the wavelet packets or the cosine packets for analyzing multivariate time series. The SLEX waveforms are complex valued and hence can be directly used to estimate the lag between components of a multivariate time series. Moreover, the SLEX waveforms are time- and frequency-localized generalizations of the Fourier waveforms. Thus, the SLEX methods parallel classical spectral analysis of stationary processes that are based on the Fourier functions. Using the SLEX library, one can develop a family of models that is a time-dependent analog of the Cramér spectral representation for stationary processes. This family of SLEX representations can be used to study time-evolving coherence and to select time–frequency spectral features for classification and discrimination of nonstationary time series.

3. Selecting the best SLEX signal representation

Our ultimate task here is to estimate the time-varying spectra and coherence of multivariate time series. We shall accomplish this by selecting *the* model, from a family of SLEX models, that best represents the data. The first step is to *build a family of SLEX models* where each of which explicitly characterizes the evolutionary spectral features of the multivariate time series and each has a spectral representation in terms of a unique SLEX basis. The second step is to *select the model* that best represents the time series data. This is equivalent to selecting the optimal dyadic segmentation of the multivariate time series. We use the penalized log energy criterion that was demonstrated in the study by Ombao et al. (2005) to be the sum of (1) the Kullback–Leibler (KL) discrepancy between a SLEX model and the unknown process that generated the time series data set and (2) the complexity penalty term that is needed to control the probability of splitting stationary time blocks. After the best model (or the best segmentation) is selected, estimates for the time-varying spectral matrix, coherence and partial coherence are extracted at the blocks that define the best segmentation.

As part of the model selection step, we address the problem of high dimensionality and multicollinearity in the multivariate time series. This problem is seen very often in brain signals. We discourage the approach of performing many separate pairwise bivariate analyses – especially when there are many channels – because this does not accurately capture how all components of the multivariate time series *simultaneously* interact with each other. Here, we promote the approach of systematically extracting a set of nonredundant spectral information that will be used in model selection and further analysis. We apply a time-varying eigenvalue–eigenvector decomposition of the SLEX spectral density matrix that generalizes the frequency-domain principal components

analysis (PCA) in the study by Brillinger (1981). The output of this step is the zero coherency (uncorrelated) SLEX principal components with time-varying spectra. The SLEX components will be utilized in the model selection step.

3.1. Building a family of SLEX models

Let $\mathbf{X}(t) = [X_1(t), \ldots, X_P(t)]'$ be a P-dimensional nonstationary time series observed on discrete time $t = 1, \ldots, T$. Our immediate goal is to find the SLEX model that best represents this time series using the complexity-penalized Kullback–Leibler criterion that is essentially a measure of divergence between the candidate SLEX model and the true unknown process that generated the observed data.

The family of SLEX models consists of signal representations, each of which uses a unique basis from the SLEX library. Thus, each model corresponds to a unique dyadic segmentation of the time series. The primary elements of each SLEX model are the following: (a) unique SLEX basis where each waveform is defined on the unit interval $[0, 1]$; (b) corresponding SLEX transfer function; and (c) zero-mean orthonormal increment random process. These elements are shared in other representations of stochastic processes such as the locally stationary processes (Dahlhaus, 2001) and locally stationary wavelet processes (Nason et al., 2000).

Let $\mathcal{B}$ be a collection of rescaled time blocks in $[0, 1]$ for one particular segmentation and denote $\{\phi_{B,\omega}(t), B \in \mathcal{B}\}$ to be one particular basis. Note that the blocks B are rescaled analogues of the blocks of time series S and also characterized by the scale or level j and shift b. Define $\Theta_B(\omega)$ as the transfer function defined on block B and $\mathrm{d}\mathbf{Z}_B(\omega)$ to be an orthonormal increment random process that satisfies

$$\mathbb{E}\mathrm{d}\mathbf{Z}_B(\omega) = 0 \tag{4}$$

$$\mathbb{C}\mathrm{ov}[\mathrm{d}\mathbf{Z}_B(\omega), \mathrm{d}\mathbf{Z}_{B'}(\lambda)] = 0 \tag{5}$$

$$\mathbb{C}\mathrm{ov}[\mathrm{d}\mathbf{Z}_B(\omega), \mathrm{d}\mathbf{Z}_B(\lambda)] = \delta(\omega - \lambda)\mathbf{1}\mathrm{d}\omega\mathrm{d}\lambda. \tag{6}$$

The SLEX model that corresponds to the segmentation $\mathcal{B}$ is

$$\mathbf{X}(t) = \sum_{B \in \mathcal{B}} \int_{-0.5}^{0.5} \Theta_B(\omega)\phi_{B,\omega}(t)\mathrm{d}\mathbf{Z}_B(\omega). \tag{7}$$

Some remarks

(i) While the time series is defined on the discrete time points $t = 1, \ldots, T$, the spectral quantities are defined on the time–frequency pair (u, ω), where u belongs to the rescaled unit interval $[0, 1]$ and ω belongs to $(-0.5, 0.5)$.

(ii) The SLEX spectral density matrix at frequency (u, ω), where u belongs to a block B is defined to be

$$\mathbf{f}(u, \omega) = \Theta_B(\omega)\Theta_B^*(\omega),$$

where Θ^* is the complex-conjugate transpose of the matrix Θ. The spectral matrix $\mathbf{f}(u, \omega)$ is a $P \times P$ Hermitian matrix.

(iii) The autospectrum of the pth component $X_p(t)$ on (u, ω) is the pth element on the diagonal denoted by $f_{pp}(u, \omega)$.
(iv) The cross-spectrum between the pth and qth components on (u, ω) is the (p, q) element $f_{pq}(u, \omega)$.
(v) The cross-coherence between the pth and qth components on (u, ω) is defined to be

$$\rho_{pq}(u, \omega) = \frac{|f_{pq}(u, \omega)|^2}{f_{pp}(u, \omega) f_{qq}(u, \omega)}.$$

3.1.1. The complexity-penalized Kullback–Leibler criterion

To select the best SLEX model, we apply the complexity-penalized Kullback–Leibler criterion that is derived in the study by Ombao et al. (2005). This criterion has two components: (1) the KL part that measures divergence between the candidate SLEX model and the process that generated the data and (2) the complexity penalty part that prevents the unnecessary splitting of a stationary mother block into children blocks. This complexity-penalized Kullback–Leibler criterion, which we simply denote as KL, explicitly takes into account both the auto- and cross-correlation information from all components of the multivariate time series simultaneously.

Consider a candidate model $\mathcal{M}_{\mathcal{B}}$ where $\mathcal{B}$ is a set of rescaled blocks $\{B(j, b)\}$ that corresponds to a particular segmentation of the time series. Let B be one block in the basis $\mathcal{B}$ and denote $\mathcal{C}(B)$ to be its corresponding KL value. The total KL for the candidate model $\mathcal{M}_{\mathcal{B}}$ is added over all blocks $\mathcal{C}(\mathcal{B}) = \sum_{B \in \mathcal{B}} C(B)$.

We state the complexity-penalty KL criterion derived in the study by Ombao et al. (2005). Consider the block of time series $\mathbf{X}(t)$, where $t \in S(j, b)$, which corresponds to the rescaled block $B(j, b)$ on the $[0, 1]$ interval. In this block, there are a total of $M_j = T/2^j$ time points and thus also a total number of M_j discrete frequency values. The KL value on block $B(j, b)$ is

$$\mathcal{C}(j, b) = \sum_{k=-M_j/2+1}^{M_j/2} \log \det \widetilde{\mathbf{I}}_{j,b}(\omega_k) + \beta_{j,b}(p)\sqrt{M_j}\,, \tag{8}$$

where $\widetilde{\mathbf{I}}_{j,b}$ is the smoothed periodogram matrix on $S(j, b)$ (see Section 2.4) and $\beta_{j,b}(p)$ is the data-driven complexity penalty for block $S_{j,b}$. Let $h_{j,b}$ be the bandwidth used in smoothing the SLEX periodogram matrix. A simple version of the complexity parameter that we will use is $\beta_{j,b}(p) = p\,\beta_{j,b}$, where $\beta_{j,b}$ takes the form

$$\beta_{j,b} = \beta_{j,b}(h_{j,b}) = \log_{10}(\mathrm{e})/\sqrt{h_{j,b}}\,\sqrt{2\,\log M_j}. \tag{9}$$

Finally, the complexity-penalized KL value for the model $\mathcal{M}_{\mathcal{B}}$ is

$$\mathcal{C}(\mathcal{B}) = \sum_{B(j,b) \in \mathcal{B}} C(j, b).$$

As an illustration, the cost for the model defined by the shaded blocks in Fig. 5 is the sum of the cost at each of these blocks: $C(1, 0) + C(2, 2) + C(2, 3)$.

3.1.2. The algorithm for selecting the best model

The best segmentation $\mathcal{B}^*$ or equivalently the best model $\mathcal{M}_{\mathcal{B}^*}$ for the data is the one that minimizes the complexity-penalized KL criterion, i.e.,

$$\mathcal{B}^* = \text{argmin}_{\mathcal{B}} \mathcal{C}(\mathcal{B}).$$

In the actual implementation, we will utilize the best basis algorithm (BBA) described in Section 2.5. This is a bottom–up algorithm and the essential idea is to compare the cost at a parent block and the children blocks. If $C(j,b) < C(j+1, 2b-1) + C(j+1, 2b)$, then we choose the parent block $S(j,b)$. Otherwise, we choose the children blocks.

3.1.3. Remarks on model selection

1. *On approximating the true process by a SLEX model.* The procedure selects, within the family of Gaussian SLEX processes, the minimizer of the Kullback–Leibler divergence between the candidate models and the true underlying process that generated the data. This is equivalent to finding the best Kullback–Leibler approximation of a piecewise stationary covariance SLEX process to the data.
2. *On extracting nonredundant information.* The components of many brain signals are usually highly collinear and consequently the estimated spectral matrix may be close to singular and hence could introduce complications in computing the KL criterion in Eq. (8). High multicollinearity in the data suggests that one should perform reduction in the dimensionality by, for example, extracting the components that give nonredundant information and account for a significant portion of the total variation in the multivariate time series. One way to accomplish this is via SLEX principal components analysis (discussed in Section 3.1.4). In the computation of the complexity-penalized KL criterion, we replace the multivariate time series by the SLEX principal components. Thus, the model selection procedure is conducted by taking into account the full information on the multivariate spectra.
3. *On the necessity of the complexity penalty term.* The penalty term is added in the criterion to prevent the method from choosing a model with unnecessarily too many blocks. For example, when the mother block is stationary, the penalty is expected to increase the probability of choosing the mother block instead of the children blocks. The form of this penalty term being proportional to $\sqrt{M_j}$, the root of the length of block $S(j,b)$, is motivated in the study by Ombao et al. (2005) and Donoho et al. (2000) to be the correct normalization following arguments that the sum $\sum_{k=-M_j/2+1}^{M_j/2} \log \det \mathbf{f}_{j,b}(\omega_k)$ can be viewed as a projection onto a Haar wavelet vector with norm $\sqrt{M_j}$. In practice, we use the smoothed periodogram matrices $\widetilde{\mathbf{I}}_{j,b}(\omega_k)$, and various simulation studies suggest that they tend to prevent the unnecessary split of stationary blocks.
4. *On the complexity penalty parameter $\beta_{j,b}$.* This will be determined from the data. The search algorithm for the best segmentation (or block partitions) can be considered as a variant of the Dyadic CART algorithm as in the study by Donoho (1997). In this algorithm, the act of excluding or including a block is analogous to Haar-wavelet thresholding of the coefficients in this block $S(j,b)$ (with

length M_j). One possibility is to use the universal threshold that is proportional to the standard deviation of these coefficients, times the well-known factor of $\sqrt{2\log M_j}$. This yields

$$\beta_{j,b} = P \times \log_{10}(e) \times \sqrt{(h_{j,b})^{-1}\ 2\ \log(M_j)},$$

where $h_{j,b}$ is the bandwidth applied when smoothing all auto- and cross-periodograms to ensure that our spectral matrix estimates are non-negative definite.

3.1.4. SLEX principal components analysis

For stationary multivariate time series, Brillinger (1981) motivates frequency-domain PCA in the following way. Let $\mathbf{X}(t)$ be a P-variate zero-mean time series with spectral density matrix $\mathbf{f}(\omega)$. Suppose now that we want to approximate $\mathbf{X}(t)$ by a $Q-$variate process ($Q \leq P$) $\mathbf{U}(t)$ whose components have zero coherency, defined to be

$$\mathbf{U}(t) = \sum_{\ell=-\infty}^{\infty} \mathbf{c}'_{t-\ell}\mathbf{V}_\ell,$$

where $\{\mathbf{c}_r\}$ is a $P \times Q$ filter matrix satisfying $\sum_{r=-\infty}^{\infty} |\mathbf{c}_r| < \infty$. We now summarize how the filter coefficients $\{\mathbf{c}_r\}$ are derived via a reconstruction criterion. Suppose that, from the reduced time series $\mathbf{U}(t)$, we want to be able to reconstruct the original time series $\mathbf{X}(t)$ by $\widehat{\mathbf{X}}(t) = \sum_{\ell=-\infty}^{\infty} \mathbf{b}_{t-\ell}\mathbf{U}(\ell)$, where the filter $\mathbf{b}_r$ is a $P \times Q$ matrix that satisfies $\sum_{r=-\infty}^{\infty} |\mathbf{b}_r| < \infty$. Here, we want $\widehat{\mathbf{X}}(t)$ to be such that the mean square approximation error $E[\big(\mathbf{X}(t) - \widehat{\mathbf{X}}(t)\big)^* \big(\mathbf{X}(t) - \widehat{\mathbf{X}}(t)\big)]$ is minimized. To simplify the discussion, suppose that the eigenvalues of $\mathbf{f}(\omega)$ are unique and we let $v^1(\omega) > v^2(\omega) > \ldots, > v^Q(\omega)$ be the eigenvalues with corresponding eigenvectors are $V^1(\omega), V^2(\omega), \ldots, V^Q(\omega)$. The solution is to choose $\mathbf{c}_\ell = \int_{-1/2}^{1/2} \mathbf{c}(\omega)\exp(i2\pi\ell\omega)\mathrm{d}\omega$, where $\mathbf{c}(\omega)$ is the matrix consisting of eigenvectors $V^1(\omega), \ldots V^Q(\omega)$. It turns out that the spectrum of the mth principal component $U_m(t)$ at frequency ω is the mth largest eigenvalue $v^m(\omega)$. For an excellent discussion on the applications of frequency-domain PCA in stationary time series, we refer the reader to the work done by Shumway and Stoffer (2006).

These ideas are extended to the nonstationary case by allowing the filter coefficients $\{\mathbf{c}_r\}$ above vary over time. Here, we decompose the multivariate nonstationary time series into the SLEX principal components, which are nonstationary components that have zero coherency. The time-varying filter and SLEX PC spectra defined on rescaled block $B(j,b)$ are obtained by performing an eigenvalue–eigenvector decomposition of the estimated spectral density matrix $\widetilde{\mathbf{I}}_{j,b}(\omega_k)$ for each ω_k on the time block $S(j,b)$. The best model (or best segmentation) is obtained by applying the penalized log energy criterion on the SLEX PC. The spectra of the SLEX PCs are simply the eigenvalues of the spectral density matrix. Denote $v^1_{j,b}(\omega_k), \ldots, v^p_{j,b}(\omega_k)$ to be the p eigenvalues arranged in decreasing magnitude. If the reduced dimension $q \leq p$ is known, then the penalized log energy criterion (8) at block $S(j,b)$ can be defined in terms of the q

SLEX PCs to be

$$\mathcal{C}(j,b) = \sum_{k=-M_j/2+1}^{M_j/2} \sum_{d=1}^{q} \log(v_{j,b}^d(\omega_k)) + q\ \beta_{j,b}\ \sqrt{M_j}. \tag{10}$$

In practice, however, q is rarely known and there is no consensus on the best approach to selecting q even in the stationary situation. We propose a data-adaptive approach that does not require the user to specify q. The basic idea is to assign a weight to each SLEX component that is proportional to its variance (spectrum). Essentially, SLEX PCs with larger eigenvalues are given more weight and those with smaller eigenvalues are given smaller weights. The weight $w_{j,b}^d(\omega_k)$ of the SLEX PC with the dth largest eigenvalue is defined to be

$$w_{j,b}^d(\omega_k) = v_{j,b}^d(\omega_k) / \sum_{c=1}^{p} v_{j,b}^c(\omega_k). \tag{11}$$

At block $S(j,b)$, the penalized log energy cost is defined to be

$$\mathcal{C}(j,b) = \sum_{k=-M_j/2+1}^{M_j/2} \sum_{d=1}^{p} w_{j,b}^d(\omega_k)\ \log v_{j,b}^d(\omega_k) + \beta_{j,b}\ \sqrt{M_j}, \tag{12}$$

where, as before, $\log(v_{j,b}^d(\omega_k))$ is the logarithm of the spectrum of the dth principal component at frequency ω_k in block $S(j,b)$ having applied PCA to the optimally smoothed periodogram matrix. Note that this cost is based on the "weighted" eigenvalues. Hence, we do not need the factor q in the complexity penalty term.

One advantage of the approach of weighting the eigenvalues is that the "optimal" number q need not be explicitly specified as it implicitly renders irrelevant to those components that do not contribute much to the variance. From a numerical point of view, it also avoids computational problems since the term $w^d \log(v^d)$ is assigned the value "zero" when v^d and w^d are both close to 0, i.e., when the absolute and relative contribution to variance, respectively, are small.

3.2. Obtaining the spectral estimates

Let $\mathcal{B}^*$ be the basis that corresponds to the best model. To estimate the time-varying spectral matrix $\mathbf{f}(u,\omega)$ at rescaled time u and frequency ω, suppose that $B(j,b)$ is the time block in the basis $\mathcal{B}^*$ that corresponds to the rescaled time u. The estimate of the SLEX spectral density matrix at (u,ω) is defined to be

$$\widehat{\mathbf{f}}(u,\omega) = \widetilde{\mathbf{I}}_{j,b}(\omega),$$

which is the kernel smoothed periodogram matrix. The autospectral estimate of the pth component $X_p(t)$ defined on (u,ω) is $\widehat{f_{pp}}(u,\omega)$; the cross-spectral estimate between

the pth and qth components is $\widehat{f}_{pq}(u,\omega)$; and the cross-coherence estimate between the pth and qth components is

$$\widehat{\rho}_{pq}(u,\omega) = \frac{|\widehat{f}_{pq}(u,\omega)|^2}{\widehat{f}_{pp}(u,\omega)\widehat{f}_{qq}(u,\omega)}.$$

To compute the confidence intervals for the SLEX autospectra, we state the asymptotic results in the study by Ombao et al. (2002). For $\omega \in (0, 1/2)$,

$$\widehat{f}_{p,p}(u,\omega)/f_{p,p}(u,\omega) \dot{\sim} \chi^2_{2M_j h_{j,b}}/(2M_j h_{j,b}),$$

where $h_{j,b}$ is the smoothing bandwidth and $M_j h_{j,b}$ is the number of frequency indices in the smoothing span. To obtain the confidence intervals for the SLEX coherence, define

$$\widehat{r}_{p,q}(u,\omega) = \tanh^{-1}[\widehat{\rho}_{p,q}(u,\omega)].$$

Then $\widehat{r}_{p,q}(u,\omega)$ is asymptotically normal with mean and variance approximately equal to $r_{p,q}(u,\omega)$ and $1/[2(2M_j h_{j,b} - P)]$, respectively. This follows readily from the study by Goodman (1963) and Brillinger (1981, Section 8.6).

3.3. An example: Multichannel EEG

The data set is an 18-channel EEG recorded from a patient of Dr. Malow (neurologist at the University of Michigan). Each EEG time series has length $T = 8192$; recorded for about 82 s and then sampled at the rate of 100 Hz. The first step in our method was to build the family of SLEX models and then select the one that is best according to our penalized log energy criterion. As suggested by the neurologist, levels $J = 6$ or $J = 7$ were used and both resulted in identical best models. Prior to model selection, the SLEX PCs were obtained via the time-varying eigenvalue–eigenvector decomposition of the SLEX matrix (see Section 3.1.4). The best model has change points that occur at approximately 20, 30, 36, 38, 40, 61, 72, 73, 74, and 77 s from the starting time. According to the neurologist, the physical manifestations of seizure became evident at around 40 s from the start of recording. However, prior to this, the SLEX analysis revealed that changes in the electrical activity of the brain were already begun to take place even before the physical symptoms were observed.

The SLEX PCA method was able to systematically filter the nonredundant information. Note that in the absence of any patient information, one would have to examine all $\frac{18!}{2!16!} = 153$ pairwise cross-correlations, which can be overwhelming. The SLEX PCA method guided the user to focus on the most interesting channels (which are $T3$ and $T4$ in this particular example). The first and second SLEX PCs altogether account for approximately 70% of the variance in the EEGs. Here, we focus our attention only to the first two SLEX PCs.

The time-varying spectra of the first SLEX PC (left side in Fig. 6) account primarily for the increase in power in the lower frequencies after the onset of seizure. The second SLEX PC (right side in Fig. 6), on the other hand, accounts for the spread of power from the delta band (0–4 Hz) to the alpha band (8–12 Hz). We further examined the magnitudes of the components of the eigenvectors at the delta, alpha, and beta

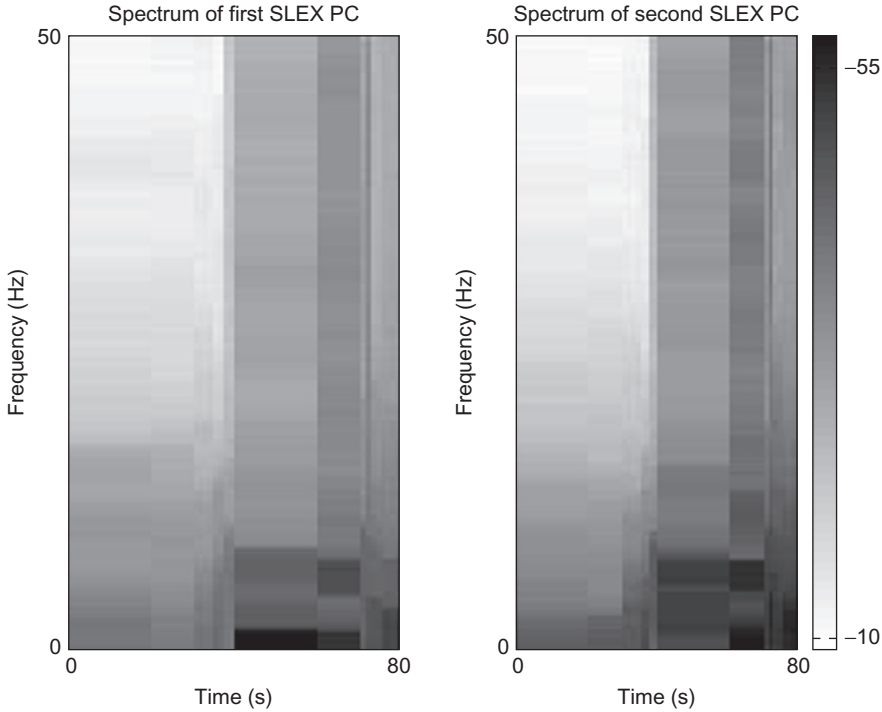

Fig. 6. Time-varying spectra of the first and second SLEX principal components.

bands (see Figs. 7 and 8), which are interpreted as the time-varying weights at the EEG channels. The SLEX method, even without using any patient information, was able to identify the $T3$ (left temporal) as one of the important channels – which is the location of a brain lesion that is believed to be implicated in seizure for this particular patient. Here, we report only the eight largest weights. We observe that, for the first SLEX PC, most of the weights are concentrated on the $T3$ (left temporal lobe) and $T4$ (right temporal lobe). For the second SLEX PC, the weights are quite diffused at the temporal and frontal lobe areas.

The estimates of the SLEX time-varying spectra at the 8 channels are in Fig. 9. The primary information that is conveyed in the spectral plots is that the distribution of power over frequency indeed evolves during the seizure process. We see that power at the lower frequencies is increased and that power is spread to middle and higher frequencies during seizure. It is important to note that features of the 18-channel EEGs were captured by the first and second SLEX PCs.

One primary goal in our analysis was to study connectivity between active brain areas, i.e., how the neuronal activity in one brain area may influence another. As suggested by the first eigenvector, two networks were examined, namely, (i) coherence between $T3$ and the other channels and (ii) coherence between $T4$ and the other channels. In Figs. 10 and 11, the connectivity between brain areas changes throughout the duration of the epileptic seizure. It is fascinating that in Fig. 10, the coherence between $T3$ and those at the left side of the brain, namely, left parietal (P3), left frontal (F3),

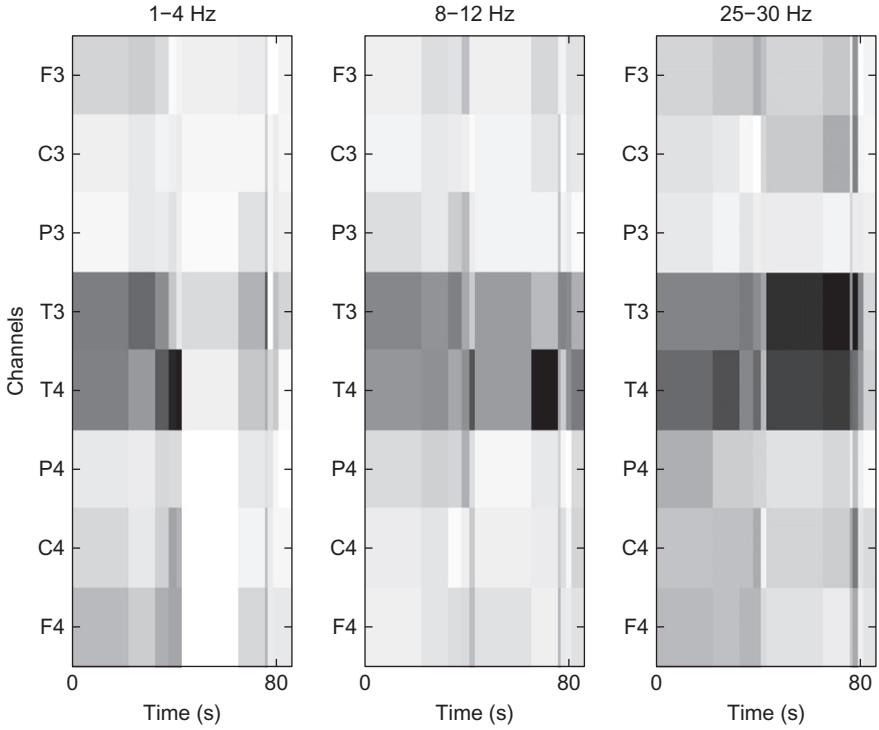

Fig. 7. Time-varying weights of the first SLEX PC at the delta frequency band (1–4 Hz), alpha frequency band (8–12 Hz), and higher beta frequency band (25–30 Hz). Darker shades represent larger weights.

and left central (C3) are very similar to each other. Moreover, coherence between T3 and the counterparts on the right side (P4, F4, and C4 respectively) are quite different – suggesting that connectivity between brain areas on the same side of the brain behave in a similar manner in the duration of the seizure. Moreover, one observes in Fig. 11 the bilaterality (symmetry) of the coherence, i.e., the coherence pattern on T4 is similar to that on T3. Bilateral synchrony is quite fascinating, though not completely well understood. It suggests rapid propagation of seizure from the left to the right temporal lobe. In addition, the existence of bilateral synchrony of the temporal lobe is not uncommon and has been observed in experimental paradigms (see the study by Grunwald et al. (1999)).

4. Classification and discrimination of time series

The extension of classical pattern-recognition techniques to nonstationary multivariate time series is a problem of great interest especially in the neuroscience community. Here, we present the SLEX method for discriminating and classifying multivariate nonstationary signals and apply this to an electroencephalogram (EEG) data set collected to study a brain network that mediates voluntary movement. In this experiment, participants performed a simple voluntary movement that required quick displacements of a

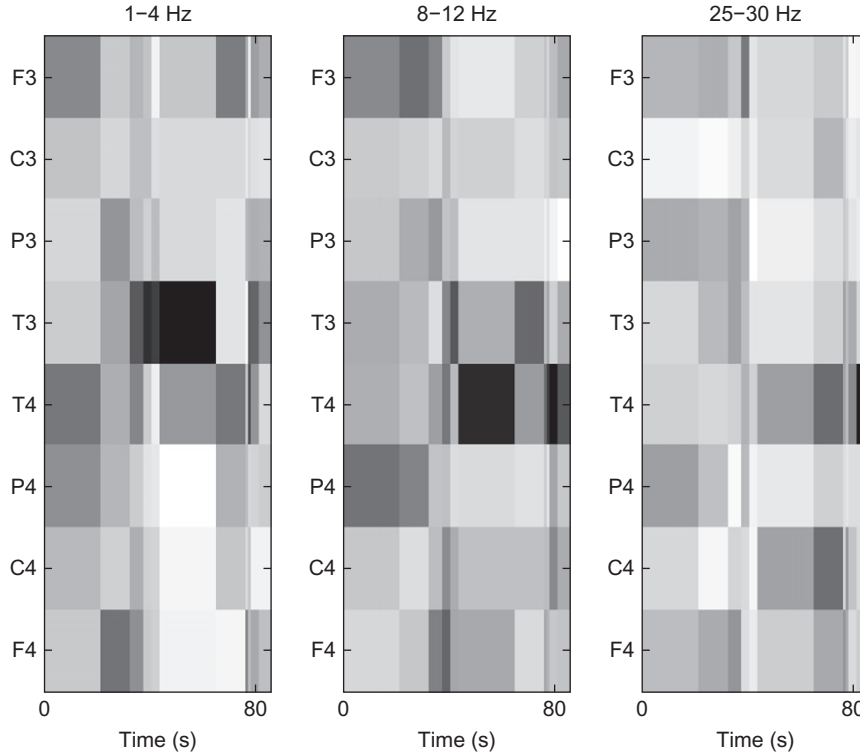

Fig. 8. Time-varying weights of the second SLEX PC at the delta frequency band (1–4 Hz), alpha frequency band (8–12 Hz), and higher beta frequency band (25–30 Hz). Darker shades represent larger weights.

hand-held joystick from a central position either to the right or to the left in response to visual cues that appeared on a computer monitor. The SLEX method is designed to *discriminate* between presumed brain connectivity occurring during leftward versus rightward movements, aiming to predict intentions to move by assessing the information evident in an electroencephalogram (EEG) time series recorded contemporary with the voluntary movements.

From a montage of 64 scalp electrodes, a set of 10 channels was preselected for further analysis (see Fig. 12). These channels were selected since they overlay regions involved in motor output (C3, C4) or visual input (O1, O2), regions that receive projections from primary neocortical visual regions (P3, P3), or regions that have involvement in action planning and have projections to neocortical motor regions (FC3, FC4, FC5, FC6). The selected sensors approximately overlay neocortical structures that have been shown to be involved in visual-motor transformations and action planning (Marconi et al., 2001). Figure 2 illustrates time-amplitude plots of the EEG obtained from a representative participant during leftward (Fig. 2, left) and rightward (Fig. 2, right) joystick movements.

Discrimination and classification of time series have a long history. Shumway and Unger (1974) and Shumway (1982) developed the framework for discrimination in time series that has been adopted in most subsequent work. Shumway and colleagues applied

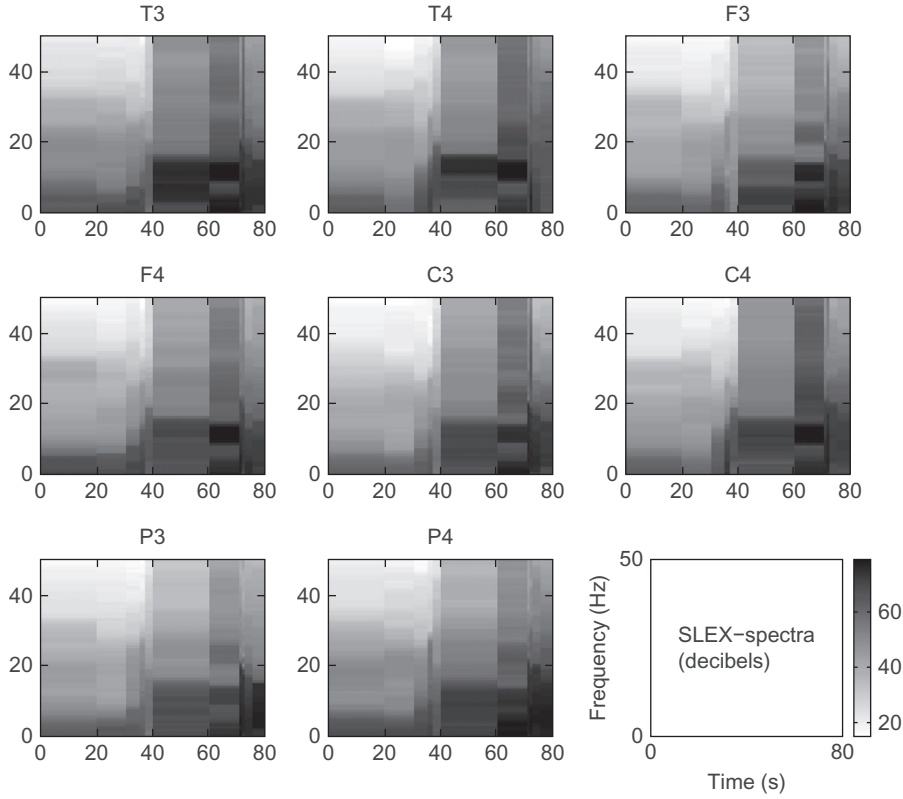

Fig. 9. SLEX autospectral estimates at the eight most important channels (chosen from the set of 18).

their work to discriminate between different seismic activities (e.g., earthquake vs. explosion). Kakizawa et al. (1998) concatenated the P-arrival and S-phases of the seismic signals into a bivariate time series and developed classification and discrimination methods for stationary multivariate time series.

For nonstationary time series, Shumway (2003) developed an information-theoretic classification method that treats the time series as realizations of the Dahlhaus (2001) model of locally stationary processes. Sakiyama and Taniguchi (2004) showed consistency of the classification procedure using the Kullback–Leibler criterion, whereas Fryzlewicz and Ombao (2009) developed a consistent classification method using stochastic wavelet representations. Saito (1994) developed another approach that selects one basis, from a collection of many bases in a library, that gives maximal separation between classes of time series.

There are a number of localized libraries that could be used for discriminating nonstationary time series. Here, we shall use the SLEX library. Huang et al. (2004), inspired by the ideas in the study by Saito (1994) and Shumway (1982), developed a procedure using the SLEX library to select the best time–frequency spectral features for discriminating between classes of univariate nonstationary time series. When extending the discrimination problem to multivariate time series, we are confronted

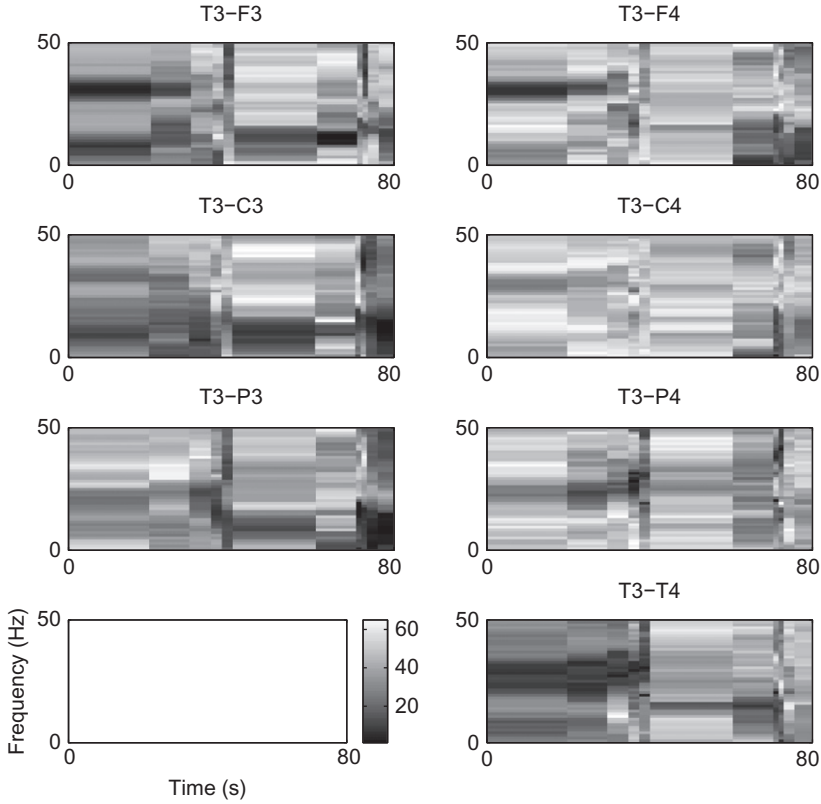

Fig. 10. Left: SLEX coherence estimates between T3 and channels on the left side of the brain namely F3, C3, P3. Right: SLEX coherence estimates between T3 and the channels on the right side of the brain namely F4, C4, P4, T4.

with two major challenges. First, most EEG datasets are massive and require computationally efficient transforms that can capture localized features of the data. Second, estimates of the multivariate spectra can be poorly conditioned (i.e., the ratio of the maximum to the minimum eigenvalue can be extremely large *leading to close-to-singular spectral estimators*). This is a consequence of the fact the sample maximum eigenvalues of a covariance (or spectral) matrix tend to over estimate the true maximum eigenvalue and the sample minimum eigenvalues tend to be negatively biased. These result in a large matrix condition number. Thus, inverting the spectral matrix estimates may give imprecise results thereby adversely impacting predictive ability especially when using information-based classification criteria such as the Chernoff criterion in Eq. (15).

A standard approach to handling highly multicollinear data entails reducing dimensionality via, for example, principal components analysis. PCA may not be ideal in discrimination and classification applications since the eigenvalue–eigenvector decomposition of the spectral matrix is invariant to (spatial) permutations of the time series. Consider a pair of channels R1 and R2 (located on the right of the scalp topography) that have a cross-dependence structure during the right-movement condition which is

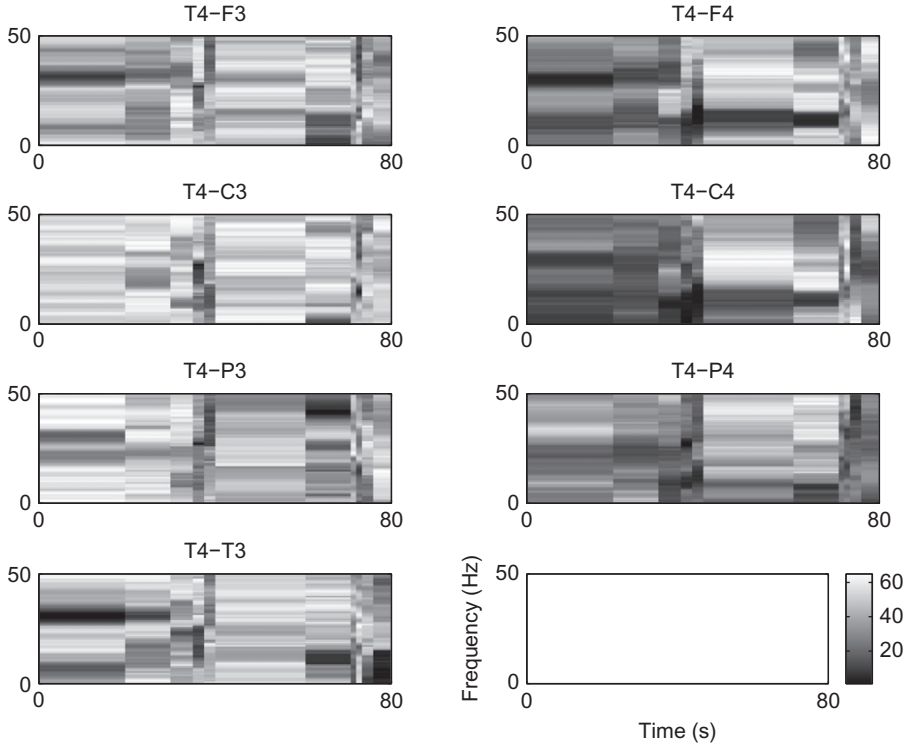

Fig. 11. Left: SLEX coherence estimates between T4 and channels on the left side of the brain namely F3, C3, P3, T3. Right: SLEX coherence estimates between T4 and the channels on the right side of the brain namely F4, C4, P4.

identical to that between a pair L1 and L2 on the left side of the scalp topography during the left movement condition. PCA is unable to distinguish the location of the sources, thereby rendering it ineffective to discriminating between the functional connectivity occurring during the leftward- and rightward-movement conditions.

Another approach to regularize the estimators is to smooth the periodogram matrices across frequency using a large enough bandwidth to prevent numerical close-to-singularity. Smoothing is a standard approach whose primary purpose is to reduce the well-known high variability of the periodogram as discussed in the study by Parzen (1961) for univariate time series and Brillinger (1981) for multivariate time series. However, unless the smoothing spans are significantly larger than the dimension P, this approach still tends to give spectral matrix estimates that are near singular. On the other hand, if the span (bandwidth) is too large, the spectral estimates will have poor frequency resolution that can dull the predictive ability, especially when the differences between conditions are present in very narrow frequency bands. Therefore, a different regularizing method that would still allow to use sufficiently small bandwidths is called for: spectral shrinkage.

Here, we present the classification and discrimination method for multivariate time series using the SLEX library to extract the localized cross-dependence structure

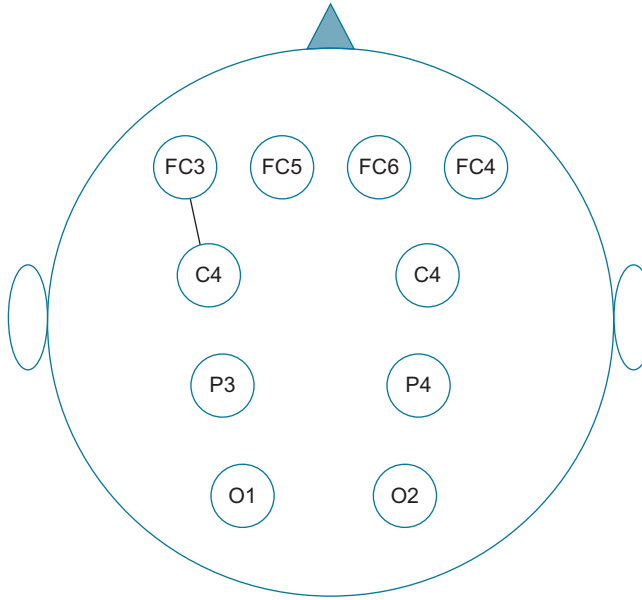

Fig. 12. The most highly discriminant network feature is the alpha band (8–12 Hz) coherence between the C3 and FC3 channels, which is significantly greater (p-value of .079 in a paired t-test) for the leftward-movement condition than the rightward-movement condition.

(brain connectivity) and the shrinkage method to estimate the spectral density matrix. This method is developed in the study by Böhm et al. (2010). The spectral shrinkage estimator is a linear combination of a mildly smoothed periodogram matrix and the identity matrix. The spectral shrinkage procedure was developed in the study by Böhm (2008) and Böhm and von Sachs (2009) for the stationary case and was further refined in the study by Fiecas et al. (2010) and Fiecas and Ombao (2011) for coherence estimation. Here, we extend this procedure to the nonstationary setting. The shrunken spectral estimator retains excellent frequency resolution, has good condition numbers, and is shown to be superior to the standard periodogram smoother in terms of the squared-error risk. Finally, as demonstrated in the simulation studies in this chapter, the shrinkage approach gives excellent classification rates.

The specific features of our approach are as follows. First, we use the SLEX library as a tool for extracting the time localized features of the nonstationary signals. Second, we estimate the time-varying spectrum via the shrinkage procedure (i.e., the slightly smoothed periodogram matrix is shrunk toward the identity matrix). In this chapter, we employ the Chernoff criterion (see Eq. (15)) that measures the divergence between the observed time series and the classes via the spectral density matrix. Furthermore, this criterion requires the computation of the inverse and the determinant of the spectral matrices. Naturally, poorly conditioned estimates result in unreliable Chernoff divergence values and, as demonstrated in this chapter, can lead to unacceptably high misclassification rates.

4.1. Overview of the shrinkage procedure for spectral estimation

Here, we summarize the basic ideas of the shrinkage procedure for spectral estimation.

4.1.1. Shrinkage for stationary time series

Let $\mathbf{X}(t) = [X_1(t), \ldots, X_P(t)]'$, $t = 1, \ldots, T$, be a stationary time series with spectral density matrix $\mathbf{f}(\omega)$. The classical estimator is the smoothed periodogram (with span m_T), which we denote as

$$\widetilde{\mathbf{f}}(\omega) = \frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \mathbf{I}(\omega + \omega_k), \quad \text{where } \omega_k = k/T.$$

Call the elements of $\widetilde{\mathbf{f}}(\omega)$ to be $\widetilde{f}_{pq}(\omega)$. Define $\widehat{\mu}_T(\omega) = \frac{1}{P} \sum_{p=1}^{P} \widetilde{f}_{pp}(\omega)$ and $\mathbf{1}$ to be the $P \times P$ identity matrix. The shrinkage estimator for $\mathbf{f}(\omega)$ takes the form

$$\widehat{\mathbf{f}}(\omega) = \frac{\widehat{\beta}_T^2(\omega)}{\widehat{\delta}_T^2(\omega)} \widehat{\mu}_T(\omega)\mathbf{1} + \frac{\widehat{\alpha}_T^2(\omega)}{\widehat{\delta}_T^2(\omega)} \widetilde{\mathbf{f}}(\omega), \tag{13}$$

where the weights are chosen as follows in order to minimize the L_2 risk in the class of considered linear combinations of $\widetilde{\mathbf{f}}(\omega)$ and the identity matrix.

First, denote $\| \mathbf{A} \|^2$ to be the *(normalized)* Hilbert-Schmidt norm of the matrix $\mathbf{A}$ (i.e., $\| \mathbf{A} \|^2 = \frac{1}{P} \operatorname{trace}(\mathbf{A}\mathbf{A}')$). Next, define

$$\widehat{\delta}_T^2(\omega) = \| \widetilde{\mathbf{f}}(\omega) - \widehat{\mu}_T(\omega)\mathbf{1} \|^2,$$

which is a measure of empirical divergence (i.e., Hilbert-Schmidt norm) between the classical smoothed periodogram and the scaled identity matrix. Define $\overline{\beta}_T^2(\omega)$ to be

$$\overline{\beta}_T^2(\omega) = \frac{1}{m_T^2} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \| \mathbf{I}(\omega + \omega_k) - \widetilde{\mathbf{f}}(\omega) \|^2,$$

which is an estimate of the local variance of the periodogram at frequency ω. Finally, $\widehat{\beta}_T^2(\omega)$ and $\widehat{\alpha}_T^2(\omega)$ are

$$\widehat{\beta}_T^2(\omega) = \min\{\overline{\beta}_T^2(\omega), \widehat{\delta}_T^2(\omega)\}$$
$$\widehat{\alpha}_T^2(\omega) = \widehat{\delta}_T^2(\omega) - \widehat{\beta}_T^2(\omega).$$

These optimal shrinkage parameters are derived using a Pythagorean relationship for their population counterparts: $\delta^2(\omega) = \alpha^2(\omega) + \beta^2(\omega)$. As shown in the study by Böhm and von Sachs (2009), the optimal shrinkage at frequency ω is the projection of the expectation of $\widetilde{\mathbf{f}}_T(\omega)$ to the line spanned by the properly scaled identity matrix and the smoothed periodogram $\widetilde{\mathbf{f}}_T(\omega)$. It is then sufficient to calculate the side lengths of the triangle built from this projection and to finally replace the population parameters $\delta^2, \alpha^2, \beta^2$ by obvious estimates using the positivity constraint.

4.1.2. Extension of shrinkage procedure for nonstationary time series

For a given nonstationary time series, the shrinkage estimator of the SLEX spectrum at rescaled block B and frequency ω_k is derived by extending the result above. In the discussion below, we shall assume that the corresponding time block for B is $S(j,b)$. Let $\mathbf{I}_{j,b}(\omega_k)$ be the SLEX periodogram at block $S(j,b)$ and frequency index k. Denote the smoothed SLEX periodogram to be

$$\widetilde{\mathbf{f}}(B,\omega_k) = \frac{1}{m_T} \sum_{\ell=-(m_T-1)/2}^{(m_T-1)/2} \mathbf{I}_{j,b}(\omega_{k+\ell})$$

and whose elements are denoted by $\widetilde{f}_{pq}(B,\omega_k)$. Denote

$$\widehat{\mu}_T(B,\omega_k) = \frac{1}{P}\sum_{p=1}^{P} \widetilde{f}_{pp}(B,\omega_k).$$

The shrinkage estimator for $\mathbf{f}(B,\omega_k)$ takes the form

$$\widehat{\mathbf{f}}(B,\omega_k) = \frac{\widehat{\beta}_T^2(B,\omega_k)}{\widehat{\delta}_T^2(B,\omega_k)}\widehat{\mu}_T(B,\omega_k)\mathbf{1} + \frac{\widehat{\alpha}_T^2(B,\omega_k)}{\widehat{\delta}_T^2(B,\omega_k)}\widetilde{\mathbf{f}}(B,\omega_k) \tag{14}$$

where the weights are derived analogously as follows:

$$\widehat{\delta}_T^2(B,\omega_k) = \| \widetilde{\mathbf{f}}(B,\omega_k) - \widehat{\mu}_T(B,\omega_k)\mathbf{1} \|^2$$
$$\widehat{\beta}_T^2(B,\omega_k) = \min\{\overline{\beta}_T^2(B,\omega_k), \widehat{\delta}_T^2(B,\omega_k)\}$$
$$\widehat{\alpha}_T^2(B,\omega_k) = \widehat{\delta}_T^2(B,\omega_k) - \widehat{\beta}_T^2(B,\omega_k),$$

where

$$\overline{\beta}_T^2(B,\omega_k) = \frac{1}{m_T^2} \sum_{\ell=-(m_T-1)/2}^{(m_T-1)/2} \| \mathbf{I}_{j,b}(\omega_{k+\ell}) - \widetilde{\mathbf{f}}(B,\omega_k) \|^2 .$$

4.2. The algorithm for the SLEX-shrinkage discrimination method

For our example, we consider a training data for each of the conditions 1 and 2 (leftward vs. rightward hand movements), which consists of P-channel time series, each having length T. There were a total of N trials for each condition and for each trial. In general, the number of trials need not be identical for the two conditions, but we make them to be so only for ease in presenting ideas. These time series from the two conditions for trial n are denoted, respectively, by

- $\mathbf{X}_n(t) = [X_{n1}(t), \ldots, X_{nP}(t)]'; n = 1, \ldots, N; t = 1, \ldots, T;$
- $\mathbf{Y}_n(t) = [Y_{n1}(t), \ldots, Y_{nP}(t)]'; n = 1, \ldots, N; t = 1, \ldots, T.$

Suppose that the data generated under these two conditions are modeled as zero-mean multivariate nonstationary processes that are characterized by their spectral matrix denoted, respectively, as $\mathbf{f}^1(u,\omega)$ and $\mathbf{f}^2(u,\omega)$. The first task is to identify the time–frequency features (autospectra, cross-spectra, coherence) that can best separate the two conditions. This is accomplished using the SLEX library as the primary tool for extracting the localized cross-dependence features and identifying the set of time blocks and frequencies that give the largest separation between $\mathbf{f}^1(u,\omega)$ and $\mathbf{f}^2(u,\omega)$. The second task is to use these selected features to classify a future signal whose group membership is not known.

The algorithm of the SLEX-shrinkage method

Consider two multivariate nonstationary processes that are characterized by the spectra denoted as $\mathbf{f}^1(u,\omega)$ and $\mathbf{f}^2(u,\omega)$, where $\mathbf{f}^g(u,\omega)$ is the $P \times P$ time-varying spectral density matrix of condition g.

Goal A: Feature extraction and selection

Step A.1 Compute the spectral matrix estimate at rescaled time block B and frequency-index k for time series in the training data.

Let $\mathbf{X}_n(t) = [X_{n1}(t), \ldots, X_{nP}(t)]'$; $n = 1, \ldots, N$; $t = 1, \ldots, T$; be the multivariate time series in the training data set recorded from N trials for condition 1. The SLEX-shrinkage spectral estimate at time block B and frequency ω_k is

$$\widehat{\mathbf{f}}^1(B,\omega_k) = \frac{1}{N}\sum_{n=1}^{N}\widehat{\mathbf{f}}_n^{\,1}(B,\omega_k),$$

where $\widehat{\mathbf{f}}_n^{\,1}(B,\omega_k)$ is the SLEX-shrinkage spectral estimate for the nth trial of condition 1. Let $\mathbf{Y}_n(t) = [Y_{n1}(t), \ldots, Y_{nP}(t)]'$; $n = 1, \ldots, N$; $t = 1, \ldots, T$; be the multivariate time series recorded from N trials for condition 2. The SLEX-shrinkage spectral estimate at time block B and frequency ω_k is denoted as $\widehat{\mathbf{f}}_n^{\,2}(B,\omega_k)$ and is computed similarly to that for condition 1.

Step A.2 Compute the *Chernoff divergence* in the spectra between the two conditions at time block B and frequency ω_k:

$$\mathcal{D}(B,\omega_k) = \ln\frac{|\lambda\widehat{\mathbf{f}}^{\,1}(B,\omega_k) + (1-\lambda)\widehat{\mathbf{f}}^{\,2}(B,\omega_k)|}{|\widehat{\mathbf{f}}^{\,2}(B,\omega_k)|} - \lambda\ln\frac{|\widehat{\mathbf{f}}^{\,1}(B,\omega_k)|}{|\widehat{\mathbf{f}}^{\,2}(B,\omega_k)|}, \quad (15)$$

where $|\mathbf{G}|$ denotes the determinant of the matrix $\mathbf{G}$ and $\lambda \in (0,1)$ is the regularization parameter. Thus, the total Chernoff divergence at time block B is

$$\mathcal{D}(B) = \sum_{k=1}^{M_B}\mathcal{D}(B,\omega_k),$$

where M_B is the number of coefficients in block B.

Step A.3 Select the most discriminant basis.
Select the best discriminant basis using the best basis algorithm outlined in Section 2.5 and denote the best basis to be the collection of blocks that we denote by $\mathcal{B}^*$.

Goal B: Classification

Consider a new time vector-valued series to be $\mathbf{Z} = [\mathbf{Z}(1), \ldots, \mathbf{Z}(T)]$ with estimated spectral matrix $\widehat{\mathbf{f}}_{\mathbf{Z}}$. The goal is to classify $\mathbf{Z}$ to the condition (either 1 or 2) to which it is least dissimilar according to the Chernoff divergence criterion. The Chernoff divergence between $\mathbf{Z}$ and conditions 1 and 2, denoted $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, is

$$\mathcal{D}_1 = \sum_{B\in\mathcal{B}^*}\sum_k \ln \frac{|\lambda\widehat{\mathbf{f}}^1(B,\omega_k) + (1-\lambda)\widehat{\mathbf{f}}_{\mathbf{Z}}(B,\omega_k)|}{|\widehat{\mathbf{f}}_{\mathbf{Z}}(B,\omega_k)|} - \lambda \ln \frac{|\widehat{\mathbf{f}}^1(B,\omega_k)|}{|\widehat{\mathbf{f}}_{\mathbf{Z}}(B,\omega_k)|}$$

$$\mathcal{D}_2 = \sum_{B\in\mathcal{B}^*}\sum_k \ln \frac{|\lambda\widehat{\mathbf{f}}^2(B,\omega_k) + (1-\lambda)\widehat{\mathbf{f}}_{\mathbf{Z}}(B,\omega_k)|}{|\widehat{\mathbf{f}}_{\mathbf{Z}}(B,\omega_k)|} - \lambda \ln \frac{|\widehat{\mathbf{f}}^2(B,\omega_k)|}{|\widehat{\mathbf{f}}_{\mathbf{Z}}(B,\omega_k)|}.$$

If $\mathcal{D}_1 > \mathcal{D}_2$, then we classify $\mathbf{Z}$ into condition 2. Otherwise, it is classified to condition 1. In our analysis, we used $\lambda = 0.50$.

4.3. Application on the visual-motor EEG data set

Electroencephalograms (EEGs) were recorded in an experiment for which five participants moved the joystick from a central position to the right when a cursor flashed on the right side of a computer monitor (or left, accordingly). There were $N = 100$ trials for each condition (right and left), and the EEG trace for each trial is a 500 ms interval with time 0 as the stimulus onset. In our analysis, we focused on the $P = 10$ channels that are believed to be most highly involved in brain motor networks engaged in visual-motor actions. These channels are (A) P3, C3, FC3, FC5, which are on the left side of the scalp topography; (B) P4, C4, FC4, FC5 on the right side; and (C) the occipital channels $O1$ and $O2$.

Our analysis showed that the best discriminant basis gives the partition $(0, 250) \cup (250, 500)$ milliseconds. This is equivalent to the segmentation $\mathcal{S}(1, 0) \cup \mathcal{S}(1, 1)$ that is the union of two halves. The difference between the right and left conditions is best captured by the partial coherence between $C3$ and $FC3$ channels at the alpha frequency band on the interval $(0, 250)$ ms, which is significantly larger in magnitude for the left condition than the right condition (see Fig. 12). This difference appears to be consistent across all five participants. We evaluated the predictive ability of the best discriminant features via a leave-one-out procedure, comparing the SLEX with shrinkage versus without shrinkage procedures. The obtained classification rates correctly identifying leftward or rightward movements are shown in the table below.
The results are very promising – shrinkage in general gives a better classification rate than nonshrinkage.

Participant	With Shrinkage (%)	Without Shrinkage (%)
1	71	65
2	72	67
3	74	66
4	74	66
5	68	71

5. Summary

We presented a systematic, flexible, and computationally efficient procedure for analyzing multivariate nonstationary time series using the SLEX library. The SLEX library is a collection of bases consisting of localized Fourier waveforms. This tool could be potentially applied to various problems in time series analysis. The SLEX waveforms are ideal for representing nonstationary time series and also for identifying time-spectral features that separate classes of time series.

Our approach to estimating the time-varying spectral features selects the basis, from the SLEX library, that best represents the time-varying auto- and cross-spectral features of the signal. The criterion for signal representation balances two important components: (a) model fit as measured by the Kullback–Leibler divergence and (b) model complexity that helps prevent unnecessary over-splitting of stationary blocks. On the problem of discrimination and classification, we select the SLEX basis that gives the maximal separation between classes as measured by the Chernoff divergence. Both the Chernoff and Kullback–Leibler divergence depend on the cross-spectral structure of the time series and thus explicitly take into account the time-varying cross-dependence between components of the time series.

These methods were illustrated for analyzing multichannel EEGs: the first application was for characterizing changes in brain electrical activity during an epileptic seizure and the second application was to discriminate between brain networks for two distinct experimental conditions.

Acknowledgments

This project is supported in part by grants from the NIH and NSF. My primary collaborators for these projects are Rainer von Sachs (Université catholique de Louvain, Belgium), Mark Fiecas (Brown University), and Hilmar Böhm (Université catholique de Louvain. Jerome N. Sanes (Neuroscience, Brown University) and Beth Malow (Neurology, Vanderbilt University) provided the EEG datasets. Daniel Van Lunen (Center for Statistical Sciences, Brown University) helped to prepare some of the figures in this chapter.

References

Auscher, P., Weiss, G.,Wickerhauser, M., 1992. Local sine and cosine basis of Coifman and Meyer and the construction of smooth wavelets. In: Chui (Ed.), Wavelets – A Tutorial in Theory and Applications. Academic Press, Boston, pp. 237–256.

Böhm, H., 2008. Shrinkage Methods for Multivariate Spectral Analysis. Ph.D. Dissertation, Université catholique de Louvain, Institut de statistique.

Böhm, H., Ombao, H., von Sachs, R., Sanes, J.N., 2010. Discrimination and classification of multivariate non-stationary signals: the SLEX-shrinkage method. J. Stat. Plan. Inference 140, 3754–3763.

Böhm, H., von Sachs, R., 2009. Shrinkage estimation in the frequency domain of multivariate time series. J. Multivar. Anal. 100, 913–935.

Brillinger, D., 1981. Time Series: Data Analysis and Theory. Holden-Day, Oakland, CA.

Coifman, R. R., Wickerhauser, M. V., 1992. Entropy-based algorithms for best basis selection. IEEE Trans. Inf. Theory 38(2), 713–718.

Dahlhaus, R., 2001. A likelihood approximation for locally stationary processes. Ann. Stat. 28, 1762–1794.

Daubechies, I., 1992. Ten Lectures on Wavelets. Society for Applied and Industrial Mathematics, Philadelphia, PA.

Donoho, D., 1997. CART and best-ortho-basis: a connection. Ann. Stat. 5, 1870–1911.

Donoho, D., Johnstone, I., 1994. Ideal adaptation via wavelet shrinkage. Biometrika 81, 425–455.

Donoho, D., Johnstone, I., 1995. Adapting to unknown smoothness via wavelet shrinkage. J. Am. Stat. Assoc. 90, 1200–1224.

Donoho, D., Mallat, S., von Sachs, R., 2000. Estimating covariances of locally stationary processes: rates of convergence of best basis methods. Technical Report 517, Department of Statistics, Stanford University.

Fiecas, M., Ombao, H., 2011. The generalized shrinkage estimator for the analysis of functional connectivity of brain signals. Ann. Appl. Stat. 5(2), 1102–1125.

Fiecas, M., Ombao, H., Linkletter, C., Thompson, W., Sanes, J.N., 2010. Functional connectivity: shrinkage estimation and randomization test. NeuroImage 40, 3005–3014.

Fryzlewicz, P., Ombao, H., 2009. Consistent classification of non-stationary signals using stochastic wavelet representations. J. Am. Stat. Assoc. 104, 299–312.

Goodman, N., 1963. Statistical analysis based upon a certain multivariate complex Gaussian distribution (an introduction). Ann. Math. Stat. 34, 152–177.

Grunwald, T., Beck, H., Lehnertz, K., Blümcke, I., Pezer, N., Kutas, M., et al., 1999. Limbic P300s in temporal lobe epilepsy with and without Ammon's horn sclerosis. Eur. J. Neurosci. 11, 1899–1906.

Huang, H.-Y., Ombao, H., Stoffer, D., 2004. Discrimination and classification of nonstationary time series using the SLEX model. J. Am. Stat. Assoc. 99, 763–774.

Kakizawa, Y., Shumway, R., Taniguchi, M., 1998. Discrimination and clustering for multivariate time series. J. Am. Stat. Assoc. 93, 328–340.

Marconi, B., Genovesio, A., Battaglia-Mayer, A., Ferraina, S., Squatrito, S., Molinari, M., et al., 2001. Eye-hand coordination during reaching. I. Anatomical relationships between parietal and frontal cortex. Cereb. Cortex 11, 513–527.

Nason, G., von Sachs, R., Kroisandt, G., 2000. Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. J. Roy. Stat. Soc. Ser. B 62, 271–292.

Ombao, H., Raz, J., von Sachs, R., Guo, W., 2002. The SLEX model of a non-stationary random process. Ann. Ins. Stat. Math. 54, 171–200.

Ombao, H., Raz, J., von Sachs, R., Malow, B., 2001. Automatic statistical analysis of bivariate nonstationary time series. J. Am. Stat. Assoc. 96, 543–560.

Ombao, H., Van Bellegem, S., 2008. Evolutionary coherence of non-stationary signals. IEEE Trans. Signal Process. 100, 101–120.

Ombao, H., von Sachs, R., Guo, W., 2005. The SLEX analysis of multivariate non-stationary time series. J. Am. Stat. Assoc. 100, 519–531.

Parzen, E., 1961. Mathematical considerations in the estimation of spectra. Technometrics 3, 167–190.

Saito, N., 1994. Local Feature Extraction and Its Applications. Ph.D. Dissertation, Yale University, Department of Mathematics.

Sakiyama, K., Taniguchi, M., 2004. Discriminant analysis for locally stationary processes. J. Multivar. Anal. 90, 282–300.

Shumway, R., 1982. Discriminant analysis for time series. In: Krishnaiah, P.R., Kanal, L.N. (Eds.), Handbook of Statistics, vol. 2. Elsevier, New York, Holland.

Shumway, R.H., 2003. Time-frequency clustering and discriminant analysis. Stat. Probab. Lett. 63, 948–956.

Shumway, R.H., Stoffer, D.S., 2006. Time Series Analysis and Its Applications with R Examples. Springer, New York.
Shumway, R.H., Unger, A.N., 1974. Linear discriminant functions for stationary time series. J. Am. Stat. Assoc. 69, 948–956.
Vidakovic, B., 1999. Statistical Modeling by Wavelets. John Wiley and Sons, New York.
Wickerhauser, M., 1994. Adapted Wavelet Analysis from Theory to Software. IEEE Press, Wellesley, MA.

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00015-6

15

An Alternative Perspective on Stochastic Coefficient Regression Models

Suhasini Subba Rao

Department of Statistics, Texas A & M University, College Station, TX 77843, USA

Abstract

The classical multiple regression model plays a very important role in statistical analysis. The typical assumption is that changes in the response variable, due to a small change in a given regressor, is constant over time. In other words, the rate of change is not influenced by any unforeseen external variables and remains the same over the entire time period of observation. This strong assumption may, sometimes, be unrealistic, for example, in areas such as social sciences, environmental sciences, etc. To account for variable dependence, the stochastic coefficient regression model was proposed, and there exists several articles that consider statistical inference for this type of model. Most of these methods use the underlying assumption of Gaussianity. In this chapter, we revisit the stochastic coefficient regression model and compare it with some statistical models that have recently been developed. We show that there is an interesting connection between stochastic coefficient regression models and locally stationary time series, which suggests that stochastic coefficient regression models can be fitted to time series whose covariance structure changes slowly over time. We consider methods of testing for randomness of the coefficients and develop parameter estimation methods that require no assumptions on the distribution of the stochastic coefficients, in particular do not require the Gaussianity assumption. Using these methods, we fit the stochastic coefficient regression model to two real data sets and their predictive performances are also examined.

Keywords: Gaussian maximum likelihood, frequency domain, locally stationary time series, multiple linear regression, nonstationarity, stochastic coefficients.

1. Introduction

The classical multiple linear regression model is ubiquitous in many fields of research. However, in situations where the response variable $\{Y_t\}$ is observed over time, it is not always possible to assume that the influence the regressors $\{x_{t,j}\}$ exert on the response Y_t is constant over time. A classical example, given by Burnett and Guthrie (1970), is when predicting air quality as a function of pollution emission. The influence the emissions have on air quality on any given day may depend on various factors such as the meteorological conditions on the current and previous days. Modeling the variable influence in a deterministic way can be too complex and a simpler method could be to treat the regression coefficients as stochastic. In order to allow for the influence of the previous regression coefficient on the current coefficient, it is often reasonable to assume that the underlying unobservable regression coefficients are stationary processes and each coefficient admits a linear process representation. In other words, a plausible model for modeling the varying influence of regressors on the response variable is

$$Y_t = \sum_{j=1}^{n} (a_j + \alpha_{t,j}) x_{t,j} + \varepsilon_t := \sum_{j=1}^{n} a_j x_{t,j} + X_t, \tag{1}$$

where $\{x_{t,j}\}$ are the deterministic regressors, $\{a_j\}$ are the mean regressor coefficients, $\mathbb{E}(X_t) = 0$ and satisfies $X_t = \sum_{j=1}^{n} \alpha_{t,j} x_{t,j} + \varepsilon_t$, $\{\varepsilon_t\}$ and $\{\alpha_{t,j}\}$ are jointly stationary linear time series with $\mathbb{E}(\alpha_{t,j}) = 0$, $\mathbb{E}(\varepsilon_t) = 0$, $\mathbb{E}(\alpha_{t,j}^2) < \infty$, and $\mathbb{E}(\varepsilon_t^2) < \infty$. We observe that this model includes the classical multiple regression model as a special case, with $\mathbb{E}(\alpha_{t,j}) = 0$ and $\mathrm{var}(\alpha_{t,j}) = 0$. The above model is often referred to as a stochastic coefficient regression (SCR) model. Such models have a long history in statistics (see Burnett and Guthrie (1970), Breusch and Pagan (1980), Duncan and Horn (1972), Fama (1977), Franke and Gründer (1995), Hildreth and Houck (1968), Newbold and Bos (1985), Pfeffermann (1984), Rosenberg (1972, 1973), Swamy (1970, 1971), Swamy and Tinsley (1980), Stoffer and Wall (1991), and Synder (1985)). For a review of this model the reader is referred to Newbold and Bos (1985). In recent years, several other statistical models have been proposed to model temporal changes; examples of such models include varying coefficient models and locally stationary processes. In this chapter, our aim is to revisit the SCR model, comparing the SCR model with these alternative models, we show that there is a close relationship between them (see Section 2). Having demonstrated that SCR models can model a wide range of time-varying behaviors, we consider methods of testing for randomness of the regression coefficients and develop parameter estimation methods that do not require any assumptions on the distribution of the stochastic coefficients.

In the aforementioned literature, it is usually assumed that $\{\alpha_{t,j}\}$ satisfies a linear process with structure specified by a finite number of parameters estimated by Gaussian maximum likelihood (GML). In the case $\{Y_t\}$ is Gaussian, the estimators are asymptotically normal and the variance of these estimators can be obtained from the inverse of the Fisher information matrix. Even in the situation $\{Y_t\}$ is non-Gaussian, the Gaussian likelihood is usually used as the objective function to be maximized, in this case the objective function is often called the quasi-Gaussian likelihood (quasi-GML).The

quasi-GML estimator is a consistent estimate of the parameters (see Ljung and Caines (1979), Caines (1988), Chapter 8.6, and Shumway and Stoffer (2006)), but when $\{Y_t\}$ is non-Gaussian, obtaining an expression for the standard errors of the quasi-GML estimators seems to be extremely difficult. Therefore, implicitly it is usually assumed that $\{Y_t\}$ is Gaussian, and most statistical inference is based on the assumption of Gaussianity. In several situations the assumption of Gaussianity may not be plausible, and there is a need for estimators that are free of distributional assumptions. In this chapter we address this issue.

In Section 3, two methods to estimate the mean regression parameters and the finite number of parameters that are characterizing the impulse response sequences of the linear processes are considered. The suggested methods are based on taking the Fourier transform of the observations, this is because spectral methods usually do not require distributional assumptions, are computationally fast, and can be analyzed asymptotically (see Dzhapharidze (1971), Dahlhaus (2000), Dunsmuir (1979), Giraitis and Robinson (2001), Hannan (1971, 1973), Shumway and Stoffer (2006), Taniguchi (1983), Whittle (1962), and Walker (1964)). Both of the proposed methods offer an alternative perspective of the SCR model and are free of any distributional assumptions. In Sections 5.2 and 5.3, we consider the asymptotic properties of the proposed estimators. A theoretical comparison of our estimators with the GML estimator, in most cases, is not possible, because it is usually not possible to obtain the asymptotic variance of the GML estimator. However, if we consider a subclass of SCR models, where the regressors are smooth, then the asymptotic variance of the GML estimator can be derived. Thus, in Section 5.4, we compare our estimator with the GML estimator for the subclass SCR models with smooth regressors, and show that both estimators have asymptotically equivalent distributions.

In Section 6 we consider two real data sets. The two time series are taken from the field of economics and environmental sciences. In the first case, the SCR model is used to examine the relationship between monthly inflation and nominal T-bills interest rates, where monthly inflation is the response and the T-bills rate is the regressor. We confirm the findings of Newbold and Bos (1985), who observe that the regression coefficient is stochastic. In the second case, we consider the influence man-made emissions (the regressors) have on particulate matter (the response variable) in Shenandoah National Park, USA. Typically, it is assumed that man-made emissions linearly influence the amount of particulate matter and a multiple linear regression model is fitted to the data. We show that there is clear evidence to suggest that the regression coefficients are random, hence the dependence between man-made emissions and particulate matter is more complicated than previously thought.

The proofs can be found in the technical report.

2. The stochastic coefficient regression model

2.1. The model

Throughout this chapter we will assume that the response variable $\{Y_t\}$ satisfies (1), where the regressors $\{x_{t,j}\}$ are observed and the following assumptions are satisfied.

ASSUMPTION 1.

(i) The stationary time series $\{\alpha_{t,j}\}$ *and* $\{\varepsilon_t\}$ *satisfy the following MA*(∞) *representations*

$$\alpha_{t,j} = \sum_{i=0}^{\infty} \psi_{i,j}\eta_{t-i,j}, \quad \text{for} \quad j = 1,\ldots,n, \quad \varepsilon_t = \sum_{i=0}^{\infty} \psi_{i,n+1}\eta_{t-i,n+1}, \tag{2}$$

where for all $1 \le j \le n+1$, $\sum_{i=0}^{\infty} |\psi_{i,j}| < \infty$, $\sum_{i=0}^{\infty} |\psi_{i,j}|^2 = 1$, $\mathbb{E}(\eta_{t,j}) = 0$, $\mathbb{E}(\eta_{t,j}^2) = \sigma_{j,0}^2 < \infty$, *for each* j, $\{\eta_{t,j}\}$ *are independent, identically distributed (i.i.d.) random variables, they are also independent over* j.

The parameters $\{\psi_{i,j}\}$ *are unknown but have a parametric form, that is there is a known function* $\psi_{i,j}(\cdot)$, *such that for some vector* $\boldsymbol{\theta}_0 = (\boldsymbol{\vartheta}_0, \boldsymbol{\Sigma}_0)$, $\psi_{i,j}(\boldsymbol{\vartheta}_0) = \psi_{i,j}$ *and* $\boldsymbol{\Sigma}_0 = diag(\sigma_{1,0}^2,\ldots,\sigma_{n+1,0}^2) = \mathrm{var}(\underline{\eta}_t)$, *where* $\underline{\eta}_t = (\eta_{t,1},\ldots,\eta_{t,n+1})$.

(ii) We define the compact parameter spaces $\Omega \subset \mathbb{R}^n$, $\Theta_1 \subset \mathbb{R}^q$ *and* $\Theta_2 \subset diag(\mathbb{R}^{n+1})$, *we have* $\sum_{i=0}^{\infty} |\psi_{i,j}(\boldsymbol{\vartheta})|^2 = 1$. *We shall assume that* $\boldsymbol{a}_0 = (a_1,\ldots,a_n)$, $\boldsymbol{\vartheta}_0$ *and* $\boldsymbol{\Sigma}_0$ *(which are defined in (i), above) lie in the interior of* Ω, Θ_1, *and* Θ_2, *respectively.*

Define the transfer function $A_j(\boldsymbol{\vartheta},\omega) = (2\pi)^{-1/2}\sum_{k=0}^{\infty}\psi_{k,j}(\boldsymbol{\vartheta})\exp(ik\omega)$, *and the spectral density* $f_j(\boldsymbol{\vartheta},\omega) = |A_j(\boldsymbol{\vartheta},\omega)|^2$. *Using this notation the spectral density of the time series* $\{\alpha_{t,j}\}$ *is* $\sigma_{j,0}^2 f_j(\boldsymbol{\vartheta}_0,\omega)$. *Let* $c_j(\boldsymbol{\theta}, t-\tau) = \sigma_j^2 \int f_j(\boldsymbol{\vartheta},\omega)\exp(i(t-\tau)\omega)d\omega$ *(hence* $cov(\alpha_{t,j},\alpha_{\tau,j}) = c_j(\boldsymbol{\theta}_0, t-\tau)$*).*

It should be noted that it is straightforward to generalize (2), such that the vector time series $\{\boldsymbol{\alpha}_t = (\alpha_{t,1},\ldots,\alpha_{t,n})\}_t$ has a vector MA(∞) representation. However, using this generalization makes the notation quite cumbersome. For this reason, we have considered the simpler case (2).

Example 1. If $\{\alpha_{t,j}\}$ and ε_t are autoregressive processes, then Assumption 1 is satisfied. That is, $\{\alpha_{t,j}\}$ and ε_t satisfy

$$\alpha_{t,j} = \sum_{k=1}^{p_j}\phi_{k,j}\alpha_{t-k,j} + \eta_{t,j} \quad j = 1,\ldots,n \quad \text{and} \quad \varepsilon_t = \sum_{k=1}^{p_{n+1}}\phi_{k,n+1}\varepsilon_{t-k} + \eta_{t,n+1},$$

where $\eta_{t,j}$ are i.i.d. random variables with $\mathbb{E}(\eta_{t,j}) = 0$ and $\mathrm{var}(\eta_{t,j}) = \sigma_j^2$ and the roots of the characteristics polynomial $1 - \sum_{k=1}^{p_j}\phi_{k,j}z^k$ lie outside unit circle. In this case, the true parameters are $\boldsymbol{\vartheta}_0 = (\phi_{1,1},\ldots,\phi_{p_{n+1},n+1})$ and $\boldsymbol{\Sigma}_0 = \mathrm{diag}\left(\frac{\sigma_1^2}{(\int g_1(\omega)d\omega)},\ldots,\frac{\sigma_{n+1}^2}{(\int g_{n+1}(\omega)d\omega)}\right)$, where $g_j(\omega) = \frac{1}{2\pi}|1 - \sum_{k=1}^{p_j}\phi_{k,j}\exp(ik\omega)|^{-2}$.

2.2. *A comparison of the SCR model with other statistical models*

In this section, we show that the SCR model is closely related to several popular statistical models. Of course, the SCR model includes the multiple linear regression model as a special case, with $\mathrm{var}(\alpha_{t,j}) = 0$ and $\mathbb{E}(\alpha_{t,j}) = 0$.

2.2.1. Varying coefficient models

In several applications, linear regression models with time-dependent parameters are fitted to the data. Examples include the varying coefficient models considered by Martinussen and Scheike (2000), where $\{Y_t\}$ satisfies

$$Y_t = \sum_{j=1}^{n} \alpha_j \left(\frac{t}{T}\right) x_{t,j} + \varepsilon_t, \quad t = 1, \ldots, T \tag{3}$$

and $\{\alpha_j(\cdot)\}$ are smooth, unknown functions and $\{\varepsilon_t\}_t$ are i.i.d. random variables with $\mathbb{E}(\varepsilon_t) = 0$ and $\text{var}(\varepsilon_t) < \infty$. Comparing this model with the SCR model, we observe that the difference between the two models lies in the modeling of the time-dependent coefficients of the regressors. In (3) the coefficient of the regressor is assumed to be deterministic, whereas the SCR model treats the coefficient as a stationary stochastic process. In some sense, one can suppose that the correlation in $\{\alpha_{t,j}\}$ determines the "smoothness" of $\{\alpha_{t,j}\}$. The higher the correlation of $\{\alpha_{t,j}\}$, the smoother the coefficients are likely to be. Thus the SCR model could be used as an alternative to varying coefficient models for modeling "rougher" changes.

2.2.2. Locally stationary time series

In this section, we show that a subclass of SCR models and the class of locally stationary linear processes defined by Dahlhaus (1996) are closely related. We restrict the regressors to be smooth, and assume there exists functions $\{x_j(\cdot)\}$ such that the regressors satisfy $x_{t,j} = x_j(\frac{t}{N})$ for some value N (setting $\frac{1}{T}\sum_t x_{t,j}^2 = 1$) and $\{Y_{t,N}\}$ satisfies

$$Y_{t,N} = \sum_{j=1}^{n} a_j x_j \left(\frac{t}{N}\right) + X_{t,N}, \quad \text{where} \quad X_{t,N} = \sum_{j=1}^{n} \alpha_{t,j} x_j \left(\frac{t}{N}\right) + \varepsilon_t \;\; t = 1, \ldots, T. \tag{4}$$

A nonstationary process can be considered locally stationary process, if in any neighborhood of t, the process can be approximated by a stationary process. We now show that $\{X_{t,N}\}$ (defined in (4)) can be considered as a locally stationary process.

PROPOSITION 1. *Suppose Assumption 1(i,ii) is satisfied, and the regressors are bounded* ($\sup_{j,v} |x_j(v)| < \infty$), *let* $X_{t,N}$ *be defined as in (4) and define the unobserved stationary process* $X_t(v) = \sum_{j=1}^{n} \alpha_{t,j} x_j(v) + \varepsilon_t$. *Then we have*

$$|X_{t,N} - X_t(v)| = O_p\left(\left|\frac{t}{N} - v\right|\right).$$

PROOF. Using the Lipschitz continuity of the regressors the proof is straightforward, hence we omit the details. □

The above result shows that in the neighborhood of t, $\{X_{t,N}\}$ can locally be approximated by a stationary process. Therefore the SCR model with slowly varying regressors can be considered as a "locally stationary" process.

We now show the converse, that is the class of locally stationary linear processes defined by Dahlhaus (1996), can be approximated to any order by an SCR model with slowly varying regressors. Dahlhaus (1996) defines locally stationary process as the stochastic process $\{X_{t,N}\}$, which satisfies the representation

$$X_{t,N} = \int A_{t,N}(\omega) \exp(it\,\omega) dZ(\omega), \tag{5}$$

where $\{Z(\omega)\}$ is a complex-valued orthogonal process on $[0, 2\pi]$ with $Z(\lambda + \pi) = Z(\lambda)$, $\mathbb{E}(Z(\lambda)) = 0$, and $\mathbb{E}\{dZ(\lambda)dZ(\mu)\} = \eta(\lambda + \nu)d\lambda d\mu$, $\eta(\lambda) = \sum_{j=-\infty}^{\infty} \delta(\lambda + 2\pi j)$ is the periodic extension of the Dirac delta function. Furthermore, there exists a Lipschitz continuous function $A(\cdot)$, such that $\sup_{\omega,t} |A(\frac{t}{N}, \omega) - A_{t,N}(\omega)| \leq KN^{-1}$, where K is a finite constant that does not depend on N.

In the following lemma we show that there always exists a SCR model that can approximate a locally stationary process to any degree.

Proposition 2. *Let us suppose that $\{X_{t,N}\}$ is a locally stationary process that satisfies (5) and $\sup_u \int |A(u, \lambda)|^2 d\lambda < \infty$. Then for any basis $\{x_j(\cdot)\}$ of $L_2[0, 1]$, and for every δ there exists an $n_\delta \in \mathbb{Z}$, such that $X_{t,N}$ can be represented as*

$$X_{t,N} = \sum_{j=1}^{n_\delta} \alpha_{t,j} x_j \left(\frac{t}{N}\right) + O_p(\delta + N^{-1}), \tag{6}$$

where $\{\boldsymbol{\alpha}_t\} = \{(\alpha_1, \ldots, \alpha_{t,n_\delta})\}_t$ is a second-order stationary vector process.

Proof. In the technical report. □

One application of the above result is that if the covariance structure of a time series is believed to change smoothly over time, then a SCR model can be fitted to the observations.

In the sections below, we will propose a method of estimating the parameters in the SCR model and use the SCR model with slowly varying parameters as a means of comparing the proposed method with existing Gaussian likelihood methods.

3. The estimators

We now consider two methods to estimate the mean regression parameters $\boldsymbol{a}_0 = (a_{1,0}, \ldots, a_{n,0})$ and the parameters $\boldsymbol{\theta}_0$ in the time series model (defined in Assumption 1).

3.1. Motivating the objective function

To motivate the objective function, consider the "localized" finite Fourier transform of $\{Y_t\}_t$ centered at t, that is $J_{Y,t}(\omega) = \frac{1}{\sqrt{2\pi m}} \sum_{k=1}^{m} Y_{t-m/2+k} \exp(ik\omega)$ (where m is even).

We can partition $J_{Y,t}(\omega)$ into the sum of deterministic and stochastic terms, $J_{Y,t}(\omega) = \sum_{j=1}^{n} a_{j,0} J_{t,m}^{(j)}(\omega) + J_{X,t}(\omega)$, where $J_{t,m}^{(j)}(\omega) = \frac{1}{\sqrt{2\pi m}} \sum_{k=1}^{m} x_{t-m/2+k,j} \exp(ik\omega)$, $J_{X,t}(\omega) = \frac{1}{\sqrt{2\pi m}} \sum_{k=1}^{m} X_{t-m/2+k} \exp(ik\omega)$, and (Y_t, X_t) are defined in (1). Let us consider the Fourier transform $J_{Y,t}(\omega)$ at the fundamental frequencies $\omega_k = \frac{2\pi k}{m}$ and define the $m(T-m)$-dimensional vectors $\mathcal{J}_{Y,T} = (J_{Y,m/2}(\omega_1), \ldots, J_{Y,T-m/2}(\omega_m))$ and $\mathcal{J}_{x,T}(\boldsymbol{a}_0) = (\sum_{j=1}^{n} a_{j,0} J_{m/2,m}^{(j)}(\omega_1), \ldots, \sum_{j=1}^{n} a_{j,0} J_{T-m/2,m}^{(j)}(\omega_m))$. In order to derive the objective function, we observe that $\mathcal{J}_{Y,T}$ is a linear transformation of the observations $\underline{Y}$, thus there exists a $m(T-m) \times T$-dimensional complex matrix, A, such that $\mathcal{J}_{Y,T} = A\underline{Y}$. Regardless of whether $\{Y_t\}$ is Gaussian or not, we treat $\mathcal{J}_{Y,T}$ as if it were a multivariate complex normal and define the quantity that is proportional to the quasi-likelihood of $\mathcal{J}_{Y,T}$ as

$$\ell_T(\boldsymbol{\theta}_0) = ((\mathcal{J}_{Y,T} - \mathcal{J}_{x,T}(\boldsymbol{a}_0))^H) \Delta_T(\boldsymbol{\theta}_0)^{-1} ((\mathcal{J}_{Y,T} - \mathcal{J}_{x,T}(\boldsymbol{a}_0)) + \log(\det \Delta_T(\boldsymbol{\theta}_0)),$$

where $\Delta_T(\boldsymbol{\theta}_0) = \mathbb{E}((\mathcal{J}_{Y,T} - \mathcal{J}_{x,T}(\boldsymbol{a}_0)(\mathcal{J}_{Y,T} - \mathcal{J}_{x,T}(\boldsymbol{a}_0))^H) = A\mathbb{E}(\underline{X}_T \underline{X}_T')A^H$ ($\underline{X}_T = (X_1, \ldots, X_T)'$ and H denotes the transpose and complex conjugate (see Picinbono (1996), Eq. (17)). Evaluating $\ell_T(\boldsymbol{\theta}_0)$ involves inverting the singular matrix $\Delta_T(\boldsymbol{\theta}_0)$. Hence it is an unsuitable criterion for estimating the parameters $\boldsymbol{a}_0$ and $\boldsymbol{\theta}_0$. Instead let us consider a related criterion, where we ignore the off-diagonal covariances in $\Delta_T(\boldsymbol{\theta}_0)$ and replace it with a diagonal matrix which shares the same diagonal as $\Delta_T(\boldsymbol{\theta}_0)$. Straightforward calculations show that when the $\Delta_T(\boldsymbol{\theta}_0)$ in $\ell_T(\boldsymbol{\theta}_0)$ is replaced by its diagonal, what remains is proportional to

$$\tilde{\ell}_T(\boldsymbol{\theta}_0) = \frac{1}{T_m m} \sum_{t=m/2}^{T-m/2} \sum_{k=1}^{m} \left(\frac{|J_{Y,t}(\omega_k) - \sum_{j=1}^{n} a_{j,0} J_{t,m}^{(j)}(\omega_k)|^2}{\mathcal{F}_{t,m}(\boldsymbol{\theta}_0, \omega_k)} + \log \mathcal{F}_{t,m}(\boldsymbol{\theta}_0, \omega_k) \right), \tag{7}$$

where $T_m = (T-m)$,

$$\begin{aligned} \mathcal{F}_{t,m}(\boldsymbol{\theta}_0, \omega) &= \frac{1}{2\pi} \sum_{r=-(m-1)}^{m-1} \exp(ir\omega) \sum_{j=1}^{n+1} c_j(\boldsymbol{\theta}_0, r) \times \frac{1}{m} \sum_{k=1}^{m-|r|} x_{t-m/2+k,j} x_{t-m/2+k+r,j} \\ &= \sum_{j=1}^{n} \sigma_{j,0}^2 \int_{-\pi}^{\pi} I_{t,m}^{(j)}(\lambda) f_j(\boldsymbol{\vartheta}_0, \omega - \lambda) d\lambda + \sigma_{n+1,0}^2 \int_{-\pi}^{\pi} I_m^{(n+1)}(\lambda) f_{n+1}(\boldsymbol{\vartheta}_0, \omega - \lambda) d\lambda, \end{aligned} \tag{8}$$

letting $x_{t,n+1} = 1$ for all t, $I_{t,m}^{(j)}(\omega) = |J_{t,m}^{(j)}(\omega)|^2$ and $I_m^{(n+1)}(\omega) = \frac{1}{2\pi m} |\sum_{k=1}^{m} \exp(ik\omega)|^2$. It is worth noting that if $\{Y_t\}$ were a second-order stationary timeseries,

then its DFT is almost uncorrelated and $\Delta_T(\boldsymbol{\theta}_0)$ would be close to a diagonal matrix (this property was used as the basis of a test for second-order stationarity by Dwivedi and Subba Rao (2011)).

3.2. Estimator 1

We use (7) to motivate the objective function of the estimator. Replacing the summand $\frac{1}{m}\sum_{k=1}^{m}$ in (7) with an integral yields the objective function

$$\mathcal{L}_T^{(m)}(\boldsymbol{a},\boldsymbol{\theta}) = \frac{1}{T_m}\sum_{t=m/2}^{T-m/2}\int_{-\pi}^{\pi}\left\{\frac{\mathcal{I}_{t,m}(\boldsymbol{a},\omega)}{\mathcal{F}_{t,m}(\boldsymbol{\theta},\omega)} + \log \mathcal{F}_{t,m}(\boldsymbol{\theta},\omega)\right\} d\omega, \tag{9}$$

where m is even and

$$\mathcal{I}_{t,m}(\boldsymbol{a},\omega) = \frac{1}{2\pi m}\left|\sum_{k=1}^{m}(Y_{t-m/2+k} - \sum_{j=1}^{n} a_j x_{t-m/2+k,j})\exp(ik\omega)\right|^2. \tag{10}$$

We recall that $\boldsymbol{\theta} = (\boldsymbol{\vartheta}, \boldsymbol{\Sigma})$, hence $\mathcal{L}_T^{(m)}(\boldsymbol{a},\boldsymbol{\theta}) = \mathcal{L}_T^{(m)}(\boldsymbol{a},\boldsymbol{\vartheta},\boldsymbol{\Sigma})$. Let $\boldsymbol{a} \in \Omega \subset \mathbb{R}^n$ and $\boldsymbol{\theta} \in \Theta_1 \otimes \Theta_2 \subset \mathbb{R}^{n+q+1}$. We use $\hat{\boldsymbol{a}}_T$ and $\hat{\boldsymbol{\theta}}_T = (\hat{\boldsymbol{\vartheta}}_T, \hat{\boldsymbol{\Sigma}}_T)$ as an estimator of $\boldsymbol{a}_0$ and $\boldsymbol{\theta}_0 = (\boldsymbol{\vartheta}_0, \boldsymbol{\Sigma}_0)$, where

$$(\hat{\boldsymbol{a}}_T, \hat{\boldsymbol{\vartheta}}_T, \hat{\boldsymbol{\Sigma}}_T) = \arg \inf_{\boldsymbol{a}\in\Omega,\boldsymbol{\vartheta}\in\Theta_1,\boldsymbol{\Sigma}\in\Theta_2} \mathcal{L}_T^{(m)}(\boldsymbol{a},\boldsymbol{\vartheta},\boldsymbol{\Sigma}). \tag{11}$$

We choose m, such that $T_m/T \to 1$ as $T \to \infty$ (thus m can be fixed, or grow at a rate slower than T).

3.3. Estimator 2

In the case that the number of regressors is relatively large, the minimization of $\mathcal{L}_T^{(m)}$ is computationally slow and has a tendency of converging to local minimums rather than the global minimum. We now suggest a second estimator that is based on estimator 1, but estimates the parameters $\boldsymbol{a}, \boldsymbol{\theta} = (\boldsymbol{\Sigma}, \boldsymbol{\vartheta})$ in two steps. Empirical studies suggest that it is less sensitive to initial values than estimator 1. In the first step of the scheme we estimate $\boldsymbol{\Sigma}_0$ and in the second step obtain an estimator of $\boldsymbol{a}_0$ and $\boldsymbol{\vartheta}_0$, thereby reducing the total number parameters to be estimated at each step. An additional advantage of estimating the variance of the coefficients in the first stage is that we can use it to determine whether a coefficient of a regressor is fixed or random.

3.3.1. The two-step parameter estimation scheme

Step 1: In the first step, we ignore the correlation of the stochastic coefficients $\{\alpha_{t,j}\}$ and errors $\{\varepsilon_t\}$ and estimate the mean regressors $\{a_{t,j}\}$ and variance of

the innovations $\boldsymbol{\Sigma}_T = \{\sigma_{j,0}^2\}$ using weighted least squares with $(\tilde{\boldsymbol{a}}_T, \tilde{\boldsymbol{\Sigma}}_T) = \arg\min_{\boldsymbol{a}\in\Omega,\boldsymbol{\Sigma}\in\Theta_2} \mathcal{L}_T(\boldsymbol{a}, \boldsymbol{\Sigma})$, where

$$\mathcal{L}_T(\boldsymbol{a}, \boldsymbol{\Sigma}) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\left(Y_t - \sum_{j=1}^{n} a_j x_{t,j}\right)^2}{\sigma_t(\boldsymbol{\Sigma})} + \log \sigma_t(\boldsymbol{\Sigma})\right) \tag{12}$$

and $\sigma_t(\boldsymbol{\Sigma}) = \sum_{j=1}^{n} \sigma_j^2 x_{t,j}^2 + \sigma_{n+1}^2$.

Step 2: We now use $\tilde{\boldsymbol{\Sigma}}_T$ to estimate $\tilde{\boldsymbol{a}}_T$ and $\boldsymbol{\vartheta}_0$. We substitute $\tilde{\boldsymbol{\Sigma}}_T$ into $\mathcal{L}_T^{(m)}$, keep $\tilde{\boldsymbol{\Sigma}}_T$ fixed and minimize $\mathcal{L}_T^{(m)}$ with respect to $(\boldsymbol{a}, \boldsymbol{\vartheta})$. We use $(\tilde{\boldsymbol{a}}_T, \tilde{\boldsymbol{\vartheta}}_T)$ as an estimator of $(\boldsymbol{a}_0, \boldsymbol{\vartheta}_0)$ where

$$(\tilde{\boldsymbol{a}}_T, \tilde{\boldsymbol{\vartheta}}_T) = \arg\min_{\boldsymbol{a}\in\Omega,\boldsymbol{\vartheta}\in\Theta_1} \mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\vartheta}, \tilde{\boldsymbol{\Sigma}}_T). \tag{13}$$

We choose m, such that $T_m/T \to 1$ as $T \to \infty$.

4. Testing for randomness of the coefficients in the SCR model

Before fitting a SCR model to the data, it is of interest to check whether there is any evidence to suggest the coefficients are random. Breusch and Pagan (1980) have proposed a test, based on the score statistic, to test the possibility that the parameters of a regression model are fixed against the alternative that they are random. Their test statistic is constructed under the assumption that the errors in the regression model are Gaussian and are identically distributed. Newbold and Bos (1985, Chapter 3) argue that the test proposed by Breusch and Pagan (1980) can be viewed as the sample correlation between the squared residuals and the regressors (under the assumption of Gaussianity). In this section, we suggest a distribution free version of the test given by Newbold and Bos (1985), to test the hypothesis that the parameters are fixed against the alternative that they are random. Further, we propose a test to test the hypothesis the parameters are random (i.i.d.) against the alternative that they are stochastic (and correlated).

To simplify notation we will consider simple regression models with just one regressor, the discussion below can be generalized to the multiple regression case. Let us consider the null hypothesis $H_0 : Y_t = a_0 + a_1 x_t + \epsilon_t$, where $\{\epsilon_t\}$ are i.i.d. random variables with $\mathbb{E}(\varepsilon_t) = 0$ and $\text{var}(\varepsilon_t) = \sigma_\epsilon^2 < \infty$ against the alternative $H_A : Y_t = a_0 + a_1 x_t + \epsilon_t$, where $\epsilon_t = \alpha_t x_t + \varepsilon_t$ and $\{\alpha_t\}$ and $\{\varepsilon_t\}$ are i.i.d. random variables with $\mathbb{E}(\alpha_t) = 0$, $\mathbb{E}(\varepsilon_t) = 0$, $\text{var}(\alpha_t) = \sigma_\alpha^2 < \infty$, and $\text{var}(\varepsilon_t) = \sigma_\varepsilon^2 < \infty$. If the alternatives were true, then $\text{var}(\epsilon_t) = x_t^2\sigma_\alpha^2 + \sigma_\varepsilon^2$, hence plotting $\text{var}(\epsilon_t)$ against x_t should give a clear positive slope. The following test is based on this observation. Suppose we observe $\{(Y_t, x_t)\}$ and use OLS to fit the model $a_0 + a_1 x_t$ to Y_t, let $\hat{\epsilon}_t$ denote the residuals. We use as the test statistic the sample correlation between $\{x_t^2\}$ and $\{\hat{\epsilon}_t^2\}$:

$$\mathcal{S}_1 = \frac{1}{T}\sum_{t=1}^{T} x_t^2\hat{\epsilon}_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T} x_t^2\right)\left(\frac{1}{T}\sum_{t=1}^{T}\hat{\epsilon}_t^2\right). \tag{14}$$

To understand how $\mathcal{S}_1$ behaves under the null and alternative, we rewrite $\mathcal{S}_1$ as

$$\mathcal{S}_1 = \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\epsilon}_t^2 - \mathbb{E}\left(\epsilon_t^2\right)\right)\left(x_t^2 - \frac{1}{T}\sum_{s=1}^{T} x_s^2\right)}_{o_p(1)} + R_1,$$

where $$R_1 = \frac{1}{T}\sum_{t=1}^{T} x_t^2\left(\mathbb{E}\left(\epsilon_t^2\right) - \frac{1}{T}\sum_{s=1}^{T}\mathbb{E}\left(\epsilon_s^2\right)\right).$$

We observe that in the case that the null is true, then $\mathbb{E}(\epsilon_t^2)$ is constant for all t and $\mathcal{S}_1 = o_p(1)$. On the other hand, when the alternative is true we have $\mathcal{S}_1 \to \mathbb{E}(R_1)$, noting that in this case

$$R_1 = \frac{1}{T}\sum_{t=1}^{T} x_t^2\sigma_\alpha^2\left(x_t^2 - \frac{1}{T}\sum_{t=1}^{T} x_s^2\right). \tag{15}$$

In the proposition below we derive the distribution of the test statistic $\mathcal{S}_1$, under the null and the alternative.

PROPOSITION 3. *Let $\mathcal{S}_1$ be defined in (14), and suppose the null is true and $\mathbb{E}(|\epsilon_t|^{4+\delta}) < \infty$ (for some $\delta > 0$) then we have $\sqrt{T}\Gamma_{1,T}^{-1/2}\mathcal{S}_1 \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$, where $\Gamma_{1,T} = \frac{\operatorname{var}(\epsilon_t^2)}{T}\sum_{t=1}^{T}\left(x_t^2 - \frac{1}{T}\sum_{s=1}^{T} x_s^2\right)^2$.*

Suppose the alternative is true and $\{\varepsilon_t\}$ and $\{\epsilon_t\}$ are i.i.d. random variables, $\mathbb{E}(|\varepsilon_t|^{4+\delta}) < \infty$ and $\mathbb{E}(|\alpha_t|^{4+\delta}) < \infty$ (for some $\delta > 0$) then we have

$$\sqrt{T}\Gamma_{2,T}^{-1/2}\left(\mathcal{S}_1 - R_1\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where R_1 is defined as in (15) and $\Gamma_{2,T} = \frac{1}{T}\sum_{t=1}^{T}\operatorname{var}((\alpha_t x_t + \epsilon_t)^2)\left(x_t^2 - \frac{1}{T}\sum_{s=1}^{T} x_s^2\right)^2$.

PROOF. In the technical report. □

We mention that in the case that the parameters in the regression model are fixed, but the variance of the errors vary over time (independent of x_t) the test statistic, $\mathcal{S}_1$, may mistakenly lead to the conclusion that the alternative were true (because in this case R_1 will be nonzero). However, if the variance varies slowly over time, it is possible to modify the test statistic $\mathcal{S}_1$ to allow for a time-dependent variance, we omit the details.

We now adapt the test above to determine whether the parameters in the regressor model are random against the alternative there is correlation. More precisely, consider the null hypothesis that the coefficients are random $H_0 : Y_t = a_0 + a_1 x_t + \epsilon_t$, where $\epsilon_t = \alpha_t x_t + \varepsilon_t$ and $\{\alpha_t\}$ are i.i.d. random variables and $\{\varepsilon_t\}$ is a stationary time series

with $\operatorname{cov}(\varepsilon_0, \varepsilon_k) = c(k)$ against the alternative that the coefficients are stochastic and correlated $H_A : Y_t = a_0 + a_1 x_t + \epsilon_t$, where $\epsilon_t = \alpha_t x_t + \varepsilon_t$ and $\{\alpha_t\}$ and $\{\varepsilon_t\}$ are stationary random variables with $\operatorname{cov}(\varepsilon_0, \varepsilon_k) = c(k)$ and $\operatorname{cov}(\epsilon_0, \epsilon_k) = \rho(k)$. We observe that if the null were true $\mathbb{E}(\epsilon_t \epsilon_{t-1}) = c(1)$, whereas if the alternative were true then $\mathbb{E}(\epsilon_t \epsilon_{t-1}) = x_t x_{t-1} \rho(1) + c(1)$, hence plotting $\epsilon_t \epsilon_{t-1}$ against $x_t x_{t-1}$ should give a clear line with a slope. Therefore, the empirical correlation between $\{\hat{\epsilon}_t \hat{\epsilon}_{t-1}\}$ and $\{x_t x_{t-1}\}$ can be used as the test statistic, and we define

$$\mathcal{S}_2 = \frac{1}{T} \sum_{t=2}^{T} x_t x_{t-1} \hat{\epsilon}_t \hat{\epsilon}_{t-1} - \left(\frac{1}{T} \sum_{t=2}^{T} x_t x_{t-1} \right) \left(\frac{1}{T} \sum_{t=2}^{T} \hat{\epsilon}_t \hat{\epsilon}_{t-1} \right). \tag{16}$$

Rewriting $\mathcal{S}_2$ we have

$$\mathcal{S}_2 = \underbrace{\frac{1}{T} \sum_{t=2}^{T} \left(\hat{\epsilon}_t \hat{\epsilon}_{t-1} - \mathbb{E}(\epsilon_t \epsilon_{t-1}) \right) \left(x_t x_{t-1} - \frac{1}{T} \sum_{s=2}^{T} x_s x_{s-1} \right)}_{o_p(1)} + R_2, \tag{17}$$

where $R_2 = \frac{1}{T} \sum_{t=2}^{T} x_t x_{t-1} \big(\mathbb{E}(\epsilon_t \epsilon_{t-1}) - \frac{1}{T} \sum_{s=2}^{T} \mathbb{E}(\epsilon_s \epsilon_{s-1}) \big)$. It is straightforward to see that if the null were true $\mathcal{S}_2 = o_p(1)$, but if the alternative were true then $\mathbb{E}(\mathcal{S}_2) \xrightarrow{\mathcal{P}} R_2$, noting that $R_2 = \frac{1}{T} \sum_{t=2}^{T} x_t x_{t-1} \big(x_t x_{t-1} \operatorname{cov}(\alpha_t, \alpha_{t-1}) - \frac{1}{T} \sum_{s=2}^{T} x_s x_{s-1} \operatorname{cov}(\alpha_s, \alpha_{s-1}) \big)$. Below we derive the distribution of $\mathcal{S}_2$ under the null and alternative.

PROPOSITION 4. *Let $\mathcal{S}_2$ be defined in (16), and suppose the null is true, that is $Y_t = a_0 + a_1 x_t + \epsilon_t$, where $\epsilon_t = \alpha_t x_t + \varepsilon_t$, and $\{\alpha_t\}$ are i.i.d. random variables $\mathbb{E}(|\alpha_t|^{4+\delta}) < \infty$ and $\{\varepsilon_t\}$ is a stationary time series which satisfies $\varepsilon_t = \sum_{j=0}^{\infty} \psi_j \eta_{t-j}$ and $\sum_j |\psi_j| < \infty$ and $\mathbb{E}(|\eta_j|^{4+\delta}) < \infty$. Then we have $\sqrt{T} \Delta_{1,T}^{-1/2} \mathcal{S}_2 \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$, where $\Delta_{1,T} = \frac{1}{T^2} \sum_{t_1,t_2=1}^{n} cov(\varepsilon_{t_1} \varepsilon_{t_1-1}, \varepsilon_{t_2} \varepsilon_{t_2-1}) v_{t_1} v_{t_2}$ and $v_t = \big(x_t^2 - \frac{1}{T} \sum_{s=2}^{T} x_s^2 \big)^2$.*

On the other hand suppose the alternative were true, that is $Y_t = a_0 + a_1 x_t + \epsilon_t$, where $\epsilon_t = \alpha_t x_t + \varepsilon_t$, and $\{\alpha_t\}$ and $\{\varepsilon_t\}$ are stationary time series, which satisfies $\varepsilon_t = \sum_{j=0}^{\infty} \psi_j \eta_{t-j}$, $\alpha_t = \sum_{j=0}^{\infty} \psi_{j,1} \eta_{t-j,1}$, $\sum_j |\psi_j| < \infty$, $\sum_j |\psi_{j,1}| < \infty$, $\mathbb{E}(|\eta_j|^8) < \infty$ and $\mathbb{E}(|\eta_{j,1}|^8) < \infty$. Then we have $\sqrt{T} \Delta_{2,T}^{-1/2} (\mathcal{S}_2 - R_2) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$, where $\Delta_{2,T} = \operatorname{var}(\mathcal{S}_2)$.

PROOF. Similar to Proposition 3. □

It is worth noting, it is not necessary to limit the test statistic to testing for correlation at lag one (as was done in $\mathcal{S}_2$). The test can be generalized to test for correlations at several lags by adapting the Portmanteau test statistic in an appropriate way.

5. Asymptotic properties of the estimators

5.1. Some assumptions

We now consider the asymptotic sampling properties of estimators 1 and 2. We need the following assumptions on the stochastic coefficients and the regressors, which we use to show consistency and the sampling distribution.

Let $|\cdot|$ denote the Euclidean norm of a vector or matrix.

ASSUMPTION 2. On the stochastic coefficients

(i) *The parameter spaces* Θ_1 *and* Θ_2 *are such that, there exists a* $\delta > 0$, *where* $\inf_{\boldsymbol{\Sigma}\in\Theta_2} \sigma_{n+1}^2 \geq \delta$ *and* $\inf_{\vartheta\in\Theta_1} \int_{-\pi}^{\pi} \left(\sum_{r=-(m-1)}^{m-1} (\frac{m-|r|}{m}) \exp(ir\lambda)\right) \cdot f_0(\vartheta, \omega - \lambda) d\lambda \geq \delta$.

(ii) *The parameter spaces* Ω, Θ_1, *and* Θ_2 *are compact.*

(iii) *The coefficients* $\psi_{i,j}$ *of the MA*(∞) *representation given in Assumption 1, satisfy* $\sup_{\vartheta\in\Theta_1} \sum_{i=0}^{\infty} |i| \cdot |\nabla_{\vartheta}^k \psi_{i,j}(\vartheta)| < \infty$ *(for all* $0 \leq k \leq 3$ *and* $1 \leq j \leq n+1$*).*

(iv) *The innovation sequences* $\{\eta_{t,j}\}$ *satisfy* $\sup_{1\leq j\leq n+1} \mathbb{E}(\eta_{t,j}^8) < \infty$.

ASSUMPTION 3. On the regressors

(i) $\sup_{t,j} |x_{t,j}| < \infty$ *and* $\frac{1}{T}\sum_{t=1}^{T} \frac{\boldsymbol{X}_t \boldsymbol{X}_t'}{\mathrm{var}(Y_t)}$ *is nonsingular for all* T $(\boldsymbol{X}_t' = (x_{t,1}, \ldots, x_{t,n}))$.

(ii) *Suppose that* $\boldsymbol{\theta}_0$ *is the true parameter. There does not exist another* $\boldsymbol{\theta}^* \in \Theta_1 \otimes \Theta_2$ *such that for all* $0 \leq r \leq m-1$ *and infinite number of* t *we have*

$$\sum_{j=1}^{n} \left(c_j(\boldsymbol{\theta}_0, r) - c_j(\boldsymbol{\theta}^*, r)\right) \sum_{k=0}^{m-|r|} x_{t-m/2+k,j} x_{t-m/2+k+r,j} = 0.$$

Let $\mathcal{J}_{T,m}(\vartheta, \omega)' = \sum_{t=m/2}^{T-m/2} \mathcal{F}_{t,m}(\vartheta, \omega)^{-1} (J_{t,m}^{(1)}(\omega), \ldots, J_{t,m}^{(n)}(\omega))$. *For all* T *and* $\vartheta \in \Theta_1$,
$\int \mathcal{J}_{T,m}(\vartheta, \omega) \mathcal{J}_{T,m}(\vartheta, \omega)' d\omega$ *is nonsingular and the smallest eigenvalue is bounded away from zero.*

(iii) *For all* T, $\mathbb{E}(\nabla_{\boldsymbol{\theta}}^2 \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0))$ *and* $\mathbb{E}(\nabla_{\boldsymbol{a}}^2 \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0))$ *are nonsingular matrices and the smallest eigenvalue is bounded away from zero.*

Assumption 2(i) ensures that $\mathcal{F}_{t,m}$ is bounded away from zero and thus $\mathbb{E}|\mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta})| < \infty$, similarly Assumption 2(iii) implies that $\sup_j \sum_{r=-\infty}^{\infty} |r \cdot \nabla^k c_j(\boldsymbol{\theta}, r)| < \infty$, therefore $\mathbb{E}|\nabla_{\boldsymbol{\theta}}^k \mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta})| < \infty$. Assumption 3(i,ii) ensures that the estimators converge to the true parameters.

5.2. Sampling properties of estimator 1: $\hat{\boldsymbol{\alpha}}_T$ and $\hat{\boldsymbol{\theta}}_T$

We first show consistency of $(\hat{\boldsymbol{a}}_T, \hat{\boldsymbol{\theta}}_T)$.

PROPOSITION 5. *Suppose Assumptions 1, 2(i,ii), and 3 are satisfied and the estimators* $\hat{\boldsymbol{a}}_T$, $\hat{\boldsymbol{\theta}}_T$ *are defined as in (11). Then we have* $\hat{\boldsymbol{a}}_T \xrightarrow{\mathcal{P}} \boldsymbol{a}_0$ *and* $\hat{\boldsymbol{\theta}}_T \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_0$, *as* $T_m \to \infty$ *and* $T \to \infty$.

PROOF. In the technical report. □

We now obtain the rate of convergence and asymptotic normality of the estimator, this requires a bound for the variance of $\mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta})$ and its derivatives. To do this we rewrite $\mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta})$ and its derivatives as a quadratic form. Substituting $J_{X,T}(\omega) = J_{Y,T}(\omega) - \sum_{j=1}^n a_j J_{t,m}^{(j)}(\omega)$ into $\mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta})$ gives

$$\begin{aligned}\mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta}) = \frac{1}{T_m}\Bigg\{ & V_T\left(\mathcal{F}_{\boldsymbol{\theta}}^{-1}\right) + 2\sum_{j=1}^n (a_{j,0} - a_j) D_T^{(j)}\left(\mathcal{F}_{\boldsymbol{\theta}}^{-1}\right) \\ & + \sum_{j_1,j_2=1}^n (a_{j_1,0} - a_{j_1})(a_{j_2,0} - a_{j_2}) H_T^{(j_1,j_2)}\left(\mathcal{F}_{\boldsymbol{\theta}}^{-1}\right) \\ & + \sum_{t=m/2}^{T-m/2} \int \log \mathcal{F}_{t,m}(\boldsymbol{\theta}, \omega) d\omega \Bigg\},\end{aligned}$$

where for a general function $\{G_t(\omega)\}_t$ we define

$$V_T(\mathcal{G}) = \sum_{t=m/2}^{T-m/2} \int \mathcal{G}_t(\omega) |J_{X,t}(\omega)|^2 d\omega = \frac{1}{2\pi m} \sum_{k=1}^m \sum_{s=k}^{T-m+k} \sum_{r=-k}^{m-k} X_s X_{s+r} g_{s+m/2-k}(r), \tag{18}$$

$$\begin{aligned}H_T^{(j_1,j_2)}(\mathcal{G}) &= \sum_{t=m/2}^{T-m/2} \int \mathcal{G}_t(\omega) J_{t,m}^{(j_1)}(\omega) J_{t,m}^{(j_2)}(-\omega) d\omega \\ &= \frac{1}{2\pi m} \sum_{k=1}^m \sum_{s=k}^{T-m+k} \sum_{r=-k}^{m-k} x_{s,j_1} x_{s+r,j_2} g_{s+m/2-k}(r), \\ D_T^{(j)}(\mathcal{G}) &= \sum_{t=m/2}^{T-m/2} \int \mathcal{G}_t(\omega) \Re\{J_{X,t}(\omega) J_{t,m}^{(j)}(-\omega)\} d\omega \\ &= \frac{1}{2\pi m} \sum_{k=1}^m \sum_{s=k}^{T-m+k} \sum_{r=-k}^{m-k} X_s x_{s+r,j} \tilde{g}_{s+m/2-k}(r),\end{aligned}$$

$g_s(r) = \int G_s(\omega) \exp(ir\omega) d\omega$ and $\tilde{g}_s(r) = \int G_s(\omega) \cos(r\omega) d\omega$. A similar expansion also holds for the derivatives of $\mathcal{L}_T^{(m)}$ (which can be used to numerically minimize

$\mathcal{L}_T^{(m)}$). Let $\nabla = \left(\frac{\partial}{\partial a_1}, \ldots, \frac{\partial}{\partial \boldsymbol{\theta}_q}\right)$ and $\nabla \mathcal{L}_T^{(m)} = (\nabla_{\boldsymbol{a}} \mathcal{L}_T^{(m)}, \nabla_\theta \mathcal{L}_T^{(m)})$, where

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta}) = \frac{1}{T_m} \Bigg\{ \Bigg(V_T(\nabla_{\boldsymbol{\theta}} \mathcal{F}^{-1}) + 2 \sum_{j=1}^{n} (a_{j,0} - a_j) D_T^{(j)}(\nabla_{\boldsymbol{\theta}} \mathcal{F}^{-1}) \Bigg) \\ + \sum_{j_1, j_2 = 1}^{n} (a_{j_1,0} - a_{j_1})(a_{j_2,0} - a_{j_2}) H_T^{(j_1, j_2)}(\nabla_{\boldsymbol{\theta}} \mathcal{F}^{-1}) \\ + \sum_{t=m/2}^{T-m/2} \int \frac{\nabla_{\boldsymbol{\theta}} \mathcal{F}_{t,m}(\boldsymbol{\theta}, \omega)}{\mathcal{F}_{t,m}(\boldsymbol{\theta}, \omega)} d\omega \Bigg\}\end{aligned}$$

$$\nabla_{a_j} \mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta}) = \frac{-2}{T_m} \left\{ D_T^{(j)}(\mathcal{F}^{-1}) + \sum_{j_1=1}^{n} (a_{j_1,0} - a_{j_1}) H_T^{(j, j_1)}(\mathcal{F}^{-1}) \right\} \tag{19}$$

and the second derivatives are

$$\nabla_{a_j} \nabla_{\boldsymbol{\theta}} \mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta}) = \frac{-2}{T_m} \left\{ D_T^{(j)}(\nabla_{\boldsymbol{\theta}} \mathcal{F}^{-1}) + \sum_{j_1=1}^{n} (a_{j_1,0} - a_{j_1}) H_T^{(j, j_1)}(\nabla_{\boldsymbol{\theta}} \mathcal{F}^{-1}) \right\}$$

$$\nabla_{a_{j_1}} \nabla_{a_{j_2}} \mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta}) = \frac{2}{T_m} H_T^{(j_1, j_2)}(\nabla_{\boldsymbol{\theta}} \mathcal{F}^{-1}) \tag{20}$$

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}^2 \mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta}) = \frac{1}{T_m} \Bigg[V_T(\nabla_{\boldsymbol{\theta}}^2 \mathcal{F}^{-1}) + 2 \sum_{j=1}^{n} (a_{j,0} - a_j) D_T^{(j)}(\nabla_{\boldsymbol{\theta}}^2 \mathcal{F}^{-1}) \\ + \sum_{j_1, j_2 = 1}^{n} (a_{j_1,0} - a_{j_1})(a_{j_2,0} - a_{j_2}) H_T^{(j_1, j_2)}(\nabla_{\boldsymbol{\theta}}^2 \mathcal{F}^{-1}) \\ + \frac{1}{T_m} \sum_{t=m/2}^{T-m/2} \Bigg\{ \int \Bigg[\frac{\nabla_{\boldsymbol{\theta}}^2 \mathcal{F}_{t,m}(\boldsymbol{\theta}, \omega)}{\mathcal{F}_{t,m}(\boldsymbol{\theta}, \omega)} \\ - \frac{\nabla_{\boldsymbol{\theta}} \mathcal{F}_{t,m}(\boldsymbol{\theta}, \omega) \nabla_{\boldsymbol{\theta}} \mathcal{F}_{t,m}(\boldsymbol{\theta}, \omega)'}{\mathcal{F}_{t,m}(\boldsymbol{\theta}, \omega)^2} \Bigg] d\omega \Bigg\} \Bigg].\end{aligned}$$

Since the Fourier coefficients in the quadratic forms of $\mathcal{L}_T^{(m)}$ and its derivatives are absolutely summable (under the stated assumptions), it can be shown that the covariance of the stochastic sequence $\{Z_{t,m}\}$, where $Z_{t,m} = \frac{1}{2\pi m} \sum_{k=1}^{m} \sum_{r=-k}^{m-k} X_t X_{t+r} g_{t+m/2-k}(r)\}_s$ is absolutely summable. This implies that the variance of $\mathcal{L}_T^{(m)}$ and its derivatives do not depend on m (so long as $m = o(T)$). We use this observation in the lemma below.

LEMMA 1. *Suppose Assumptions 1, 2(i–iii) (and* $\sup_j \mathbb{E}(\eta_{t,j}^4) < \infty$*), and 3(i) are satisfied. Let* $V_T(\mathcal{G})$ *and* $D_T^{(j)}(\mathcal{G})$ *be defined as in (18), where* $\sup_s \sum_r |g_s(r)| < \infty$ *and* $\sup_s \sum_r |\tilde{g}_s(r)| < \infty$*. Then we have* $\mathbb{E}(D_T^{(j)}(\mathcal{G})) = 0$*,*

$$\mathbb{E}(V_T(\mathcal{G})) = \sum_{t=m/2}^{T-m/2} \int \mathcal{G}_t(\omega)\mathcal{F}_{t,m}(\boldsymbol{\theta}_0, \omega)d\omega, \tag{21}$$

$$\operatorname{var}(V_T(\mathcal{G})) \leq T(n+1) \sup_s \sum_r |g_s(r)| \left\{ 2\left(\sum_r \rho_2(r)\right)^2 + \sum_{k_1,k_2,k_3} \rho_4(k_1,k_2,k_3) \right\} \tag{22}$$

and

$$\operatorname{var}\left(D_T^{(j)}(\mathcal{G})\right) \leq T(n+1) \sup_s \left(\sum_r |\tilde{g}_s(r)|\right)\left(\sum_r \rho_2(r)\right), \tag{23}$$

where

$$\begin{aligned} \rho_2(k) &= \kappa_2^2 \sup_j \sum_i |\psi_{i,j}| \cdot |\psi_{i+k,j}|, \\ \rho_4(k_1,k_2,k_3) &= \kappa_4 \sup_j \sum_i |\psi_{i,j}| \cdot |\psi_{i+k_1,j}| \cdot |\psi_{i+k_2,j}| \cdot |\psi_{i+k_3,j}|, \end{aligned} \tag{24}$$

$\kappa_2 = \sup_j \operatorname{var}(\eta_{0,j})$ *and* $\kappa_4 = \sup_j \operatorname{cum}(\eta_{0,j}, \eta_{0,j}, \eta_{0,j}, \eta_{0,j})$.

PROOF. See the technical report. □

Applying the mean value theorem pointwise to $\nabla \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0)$ and using the uniform convergence $\sup_{\boldsymbol{a},\boldsymbol{\theta}} |\nabla^2 \mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta}) - \mathbb{E}(\nabla^2 \mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta}))| \xrightarrow{\mathcal{P}} 0$, we have

$$\begin{pmatrix} \hat{\boldsymbol{a}}_T - \boldsymbol{a}_0 \\ \hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0 \end{pmatrix} = \mathbb{E}\big(\nabla^2 \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0)\big)^{-1} \begin{pmatrix} \nabla_{\boldsymbol{a}} \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0) \\ \nabla_{\boldsymbol{\theta}} \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0) \end{pmatrix} + o_p\left(\frac{1}{\sqrt{T}}\right). \tag{25}$$

Thus by using the above, if Assumption 3(iii) holds and $m = o(T)$, then we have

$$(\hat{\boldsymbol{a}}_T - \boldsymbol{a}_0, \hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) = O_p\left(\frac{1}{\sqrt{T}}\right).$$

In the following theorem, we use the above expansion to show asymptotic normality of $(\hat{\boldsymbol{a}}_T, \hat{\boldsymbol{\theta}}_T)$. This requires us to evaluate $\mathbb{E}(\nabla^2 \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0))$ and $\operatorname{var}(\nabla \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0))$. The

submatrices of $\mathbb{E}(\nabla^2 \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0))$ are $\mathbb{E}(\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{a}} \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0) = 0$,

$$\mathbb{E}(\nabla_{\boldsymbol{a}}^2 \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0))\rfloor_{j_1,j_2} = \frac{1}{T_m} \sum_{t=m/2}^{T-m/2} \int \mathcal{F}_{t,m}(\boldsymbol{\theta}_0, \omega)^{-1} J_{t,m}^{(j_1)}(\omega) J_{t,m}^{(j_2)}(-\omega) d\omega, \tag{26}$$

$$\mathbb{E}(\nabla_{\boldsymbol{\theta}}^2 \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0)) = \frac{1}{T_m} \sum_{t=m/2}^{T-m/2} \int \frac{\nabla \mathcal{F}_{t,m}(\boldsymbol{\theta}_0, \omega) \nabla \mathcal{F}_{t,m}(\boldsymbol{\theta}_0, \omega)'}{(\mathcal{F}_{t,m}(\boldsymbol{\theta}_0, \omega))^2} d\omega,$$

In order to compare the limiting distributions of estimator 1 and 2, we rewrite $\operatorname{var}(\nabla \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0))$ in terms of a block matrix

$$\begin{aligned} W_T &= T \operatorname{var}(\nabla \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0)) \\ &= \begin{pmatrix} W_{1,T} & T \operatorname{cov}(\nabla_{\boldsymbol{\beta}} \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0), \nabla_{\Sigma} \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0)) \\ T \operatorname{cov}(\nabla_{\boldsymbol{\beta}} \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0), \nabla_{\Sigma} \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0)) & T \operatorname{var}(\nabla_{\boldsymbol{\Sigma}} \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0)) \end{pmatrix} \end{aligned} \tag{27}$$

with $\boldsymbol{\beta} = (\boldsymbol{a}, \boldsymbol{\vartheta})$ and

$$W_{1,T} = \operatorname{var} \begin{pmatrix} \sqrt{T} \nabla_{\boldsymbol{a}} \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\vartheta}_0, \boldsymbol{\sigma}_0) \\ \sqrt{T} \nabla_{\boldsymbol{\vartheta}} \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\vartheta}_0, \boldsymbol{\sigma}_0) \end{pmatrix}. \tag{28}$$

THEOREM 1. *Suppose Assumptions 1, 2, and 3 hold, then we have*

$$\sqrt{T} B_T^{-1/2} \begin{pmatrix} \hat{\boldsymbol{a}}_T - \boldsymbol{a}_0 \\ \hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0 \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, I), \tag{29}$$

as $T_m/T \to 1$ and $T \to \infty$, where $B_T = V_T^{-1} W_T V_T^{-1}$, W_T is defined in (27) and

$$V_T = \begin{pmatrix} \mathbb{E}(\nabla_{\boldsymbol{a}}^2 \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0)) & 0 \\ 0 & \mathbb{E}(\nabla_{\boldsymbol{\theta}}^2 \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0)) \end{pmatrix}.$$

PROOF. In the technical report. □

We observe that V_T is a block diagonal matrix, this is because straightforward calculations show $\mathbb{E}(\nabla_{\boldsymbol{a}} \nabla_{\boldsymbol{\theta}} \mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\theta}_0)) = 0$.

REMARK 1. Examining W_T we observe that the only parameters in W_T, which we need to estimate (in addition to $(\boldsymbol{a}_0, \boldsymbol{\theta}_0)$) are the cumulants $\operatorname{cum}(\eta_{t,j}, \eta_{t,j}, \eta_{t,j})$ and $\operatorname{cum}(\eta_{t,j}, \eta_{t,j}, \eta_{t,j}, \eta_{t,j})$. We estimate the cumulants using the estimated moments. To estimate the moments, we group the observations $\{Y_t\}$ in $(n+1)$ blocks, each of

length $M = T/(n+1)$, and evaluate the empirical third moment within each block. If the size of each block $M = T/(n+1)$ is large, we obtain the following approximate equations

$$\frac{1}{M}\sum_{s=1}^{M} Y_{Mr+s}^3 \approx \frac{1}{M}\sum_{j=1}^{n+1}\mathbb{E}(\eta_{t,j}^3)\sum_{s=1}^{M} x_{Mr+s,j}^3 \sum_{i=0}^{\infty}\psi_{i,j}(\boldsymbol{\vartheta}_0)^3.$$

Obviously, this equation is true for $r = 1, \ldots, (n+1)$. Therefore, we have $(n+1)$ linear simultaneous equations in the unknown $\{\mathbb{E}(\eta_{t,j}^3)\}$, which, if we replace $\boldsymbol{\vartheta}_0$ with $\hat{\boldsymbol{\vartheta}}_T$, we can solve for. Thus we have an estimator of $\mathbb{E}(\eta_{t,j}^3)$. Using a similar method, the empirical fourth moment can be used to obtain an estimator of the fourth-order cumulants.

5.3. *Sampling properties of estimator 2:* $\tilde{\boldsymbol{\Sigma}}_T$, $\tilde{\boldsymbol{a}}_T$, $\tilde{\boldsymbol{\vartheta}}_T$

We now obtain the sampling properties of estimator 2. We first consider the properties of the variance estimator $\tilde{\boldsymbol{\Sigma}}_T$.

PROPOSITION 6. *Suppose Assumptions 1, 2, and 3 are satisfied, let* $\mathcal{L}_T(\boldsymbol{a}, \boldsymbol{\Sigma})$ *and* $\tilde{\boldsymbol{\Sigma}}_T$ *be defined as in (12). Then we have*

$$\begin{aligned} &\operatorname{var}(\nabla_{\boldsymbol{\Sigma}}\mathcal{L}_T(\boldsymbol{a}_0, \boldsymbol{\Sigma}_0))^{-1/2}\nabla_{\boldsymbol{\Sigma}}\mathcal{L}_T(\boldsymbol{a}_0, \boldsymbol{\Sigma}_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I) \\ &C_T^{-1/2}(\tilde{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I), \end{aligned} \tag{30}$$

as $T \to \infty$, *where* $C_T = \mathbb{E}\big(\nabla_{\boldsymbol{\Sigma}}^2\mathcal{L}_T(\boldsymbol{a}_0, \boldsymbol{\Sigma}_0)\big)^{-1}\operatorname{var}(\nabla_{\boldsymbol{\Sigma}}\mathcal{L}_T(\boldsymbol{a}_0, \boldsymbol{\Sigma}_0))\mathbb{E}\big(\nabla_{\boldsymbol{\Sigma}}^2\mathcal{L}_T(\boldsymbol{a}_0, \boldsymbol{\Sigma}_0)\big)^{-1}$.

We now consider the properties of $(\tilde{\boldsymbol{a}}_T, \tilde{\boldsymbol{\vartheta}}_T)$, which are obtained in step 2 of estimator 2.

THEOREM 2. *Suppose Assumptions 1, 2, and 3 hold, then we have*

$$\sqrt{T}(\tilde{V}_T^{(m)})^{1/2}(\tilde{W}_T^{(m)})^{-1/2}(\tilde{V}_T^{(m)})^{1/2}\begin{pmatrix}\tilde{\boldsymbol{a}}_T - \boldsymbol{a}_0 \\ \tilde{\boldsymbol{\vartheta}}_T - \boldsymbol{\vartheta}_0\end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, I), \tag{31}$$

for $m = o(T)$ *as* $T \to \infty$, $C_T = \tilde{V}_T^{-1}\tilde{W}_T\tilde{V}_T^{-1}$, *where*

$$\begin{aligned} \tilde{W}_T &= W_{T,1} + \begin{pmatrix} 0 & \Xi_1 \\ \Xi_1' & \Xi_2 \end{pmatrix}, \\ \tilde{V}_T^{(m)} &= \begin{pmatrix} \mathbb{E}(\nabla_{\boldsymbol{a}}^2\mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\vartheta}_0, \boldsymbol{\Sigma}_0)) & 0 \\ 0 & \mathbb{E}(\nabla_{\boldsymbol{\vartheta}}^2\mathcal{L}_T^{(m)}(\boldsymbol{a}_0, \boldsymbol{\vartheta}_0, \boldsymbol{\Sigma}_0)) \end{pmatrix} \end{aligned}$$

with $W_{T,1}$ defined as in (28) and

$$\Xi_1 = cov\Big(\sqrt{T}\nabla_{\boldsymbol{a}}\mathcal{L}_T^{(m)}(\boldsymbol{a}_0,\boldsymbol{\vartheta}_0,\boldsymbol{\Sigma}_0),\sqrt{T}\nabla_{\boldsymbol{\Sigma}}\mathcal{L}_T(\boldsymbol{a}_0,\boldsymbol{\Sigma}_0)\Big)\mathcal{Q}_T'$$

$$\Xi_2 = 2cov\Big(\sqrt{T}\nabla_{\boldsymbol{\vartheta}}\mathcal{L}_T^{(m)}(\boldsymbol{a},\boldsymbol{\vartheta}_0,\boldsymbol{\Sigma}_0),\sqrt{T}\nabla_{\boldsymbol{\Sigma}}\mathcal{L}_T(\boldsymbol{a}_0,\boldsymbol{\Sigma}_0)\Big)\mathcal{Q}_T' + \mathcal{Q}_T\mathrm{var}\big(\sqrt{T}\nabla_{\boldsymbol{\Sigma}}\mathcal{L}_T(\boldsymbol{a}_0,\boldsymbol{\Sigma}_0)\big)\mathcal{Q}_T',$$

Q_T is a $q \times (n+1)$-dimensional matrix defined by

$$\mathcal{Q}_T = \left(\frac{1}{T_m}\int\sum_{t=m/2}^{T-m/2}\big(\nabla_{\boldsymbol{\vartheta}}\mathcal{F}_{t,m}(\boldsymbol{\vartheta}_0,\boldsymbol{\Sigma}_0,\omega)^{-1}\big)\otimes\underline{H}_{(t,m)}(\omega)d\omega\right) \times \mathbb{E}\left(\nabla_{\boldsymbol{\Sigma}}^2\mathcal{L}_T(\boldsymbol{a}_0,\boldsymbol{\Sigma}_0)\right)^{-1} \tag{32}$$

$$h_j^{(t,m)}(\omega) = \int_{-\pi}^{\pi} I_{t,m}^{(j)}(\lambda) f_j(\boldsymbol{\vartheta}_0,\omega-\lambda)d\lambda,$$

$\underline{H}_{(t,m)}(\omega) = (h_1^{(t,m)}(\omega),\ldots,h_{n+1}^{(t,m)}(\omega))$ and noting that $\otimes$ denotes the tensor product.

REMARK 2. It is unclear which of the two estimators have the smallest variance. However, comparing the variances of the two estimators in (29) and (31), we observe that they are similar. In particular $\tilde{V}_T$ is a submatrix of V_T. The terms Ξ_1 and Ξ_2 in $\tilde{W}_T$, are due to the estimation of $\boldsymbol{\Sigma}_0$ in the first stage of the scheme. □

5.4. The Gaussian likelihood and asymptotic efficiency of estimator 1

In this section, we compare the asymptotic properties of the frequency domain estimator $(\hat{\boldsymbol{a}}_T,\hat{\boldsymbol{\theta}}_T)$ with the Gaussian maximum likelihood estimator (GMLE). We recall the GMLE is constructed as if the stochastic coefficients $\{\alpha_{t,j}\}$ and errors $\{\varepsilon_t\}$ were Gaussian. However, unlike the frequency domain estimators, in general there does not exist an explicit expression for the asymptotic variance of the Gaussian maximum likelihood. Instead we will consider a subclass of SCR models, where the regressors vary slowly over time and do the comparison over this subclass. We will show that for this subclass an asymptotic expression for the asymptotic distributional variance of the GMLE can be derived. We will assume that the regressors are such that there exists a "smooth" function, $x_j(\cdot)$, such that $x_{t,j} = x_j(\frac{t}{N})$ and $Y_t := Y_{t,N}$ satisfies

$$Y_{t,N} = \sum_{j=1}^{n}(a_{j,0}+\alpha_{t,j})x_j\left(\frac{t}{N}\right)+\varepsilon_t \quad t=1,\ldots,T. \tag{33}$$

In the following lemma, we obtain the asymptotic distribution of the GMLE under the asymptotic framework that both T and $N \to \infty$.

LEMMA 2. *Let us suppose that* $\{Y_{t,N}\}$ *satisfies (33), where the* $\{\alpha_{t,j}\}$ *and* $\{\varepsilon_t\}$ *are Gaussian and satisfy Assumption 1. Let*

$$\mathcal{F}(v,\boldsymbol{\theta}_0,\omega) = \sum_{j=1}^{n} x_j(v)^2\sigma_{j,0}^2 f_j(\boldsymbol{\vartheta}_0,\omega) + \sigma_{n+1,0}^2 f_{n+1}(\boldsymbol{\vartheta}_0,\omega). \tag{34}$$

We assume that there does not exist another $\boldsymbol{\theta} \in \Theta_1 \otimes \Theta_2$ *such that* $\mathcal{F}(v,\boldsymbol{\theta}_0,\omega) = \mathcal{F}(v,\boldsymbol{\theta}_0,\omega)$ *for all* $v \in [0,T/N]$ *and the matrix* $\frac{N}{T}\int_0^{T/N} \mathbf{x}(v)\mathbf{x}(v)'dv$, *(with* $\mathbf{x}(v)' = (x_1(v),\ldots,x_n(v))$*) has eigenvalues which are bounded from above and away from zero. Suppose* $(\boldsymbol{a}_{mle},\boldsymbol{\theta}_{mle})$ *is the Gaussian maximum likelihood estimator of the parameters* $(\boldsymbol{a}_0,\boldsymbol{\theta}_0)$. *Then we have*

$$\sqrt{T}\begin{pmatrix}\boldsymbol{a}_{mle}-\boldsymbol{a}_0\\ \boldsymbol{\theta}_{mle}-\boldsymbol{\theta}_0\end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0,\begin{pmatrix}\Delta_1^{-1} & 0\\ 0 & \Delta_2^{-1}\end{pmatrix}\right)$$

with $N \to \infty$ *as* $T \to \infty$, *where*

$$(\Delta_1)_{j_1,j_2} = \frac{N}{T}\int_0^{T/N} x_{j_1}(v)x_{j_2}(v)\mathcal{F}(v,\boldsymbol{\theta}_0,0)^{-1}dv,$$

$$\Delta_2 = 2\frac{N}{T}\int_0^{T/N}\int_0^{2\pi} \frac{\nabla_{\boldsymbol{\theta}}\mathcal{F}(v,\boldsymbol{\theta}_0,\omega)(\nabla_{\boldsymbol{\theta}}\mathcal{F}(v,\boldsymbol{\theta}_0,-\omega))'}{|\mathcal{F}(v,\boldsymbol{\theta}_0,\omega)|^2}d\omega dv,$$

$$a(v,k) = \int \frac{1}{\mathcal{F}(v,\boldsymbol{\theta}_0,\omega)}\exp(ik\omega)d\omega \underline{b}(v,k) = \int \nabla_{\boldsymbol{\theta}}\mathcal{F}(v,\boldsymbol{\theta}_0,\omega)^{-1}\exp(ik\omega)d\omega$$

$a(v) = \{a(v,k)\}$ *and* $\bar{\underline{b}}(v) = \{\underline{b}(v,-k)\}$.

In practice, for any given set of regressors $\{x_{t,j}\}$, N will not be known, but a lower bound for N can be obtained from $\{x_{t,j}\}$. To ensure the magnitude of the regressors do not influence N, we will assume that the regressors satisfy $\frac{1}{T}\sum_{t=1}^{T} x_{t,j}^2 = 1$ (for all j). To measure the smoothness of the regressors define

$$\hat{N} = \frac{1}{\sup_{t,j}|x_{t,j} - x_{t-1,j}|}. \tag{35}$$

Clearly if $\hat{N}$ is large, this indicates that the regressors are smooth.

We now compare the asymptotic variance of the GMLE and estimator 1.

PROPOSITION 7. *Suppose Assumptions 1, 2, and 3 hold and*

$$\sup_j \int \left|\frac{d^2 f_j(\boldsymbol{\vartheta}_0,\omega)}{d\omega^2}\right|^2 d\omega < \infty \quad \text{and} \quad \sup_j \int \left|\frac{d^2\nabla_{\boldsymbol{\vartheta}} f_j(\boldsymbol{\vartheta}_0,\omega)}{d\omega^2}\right|^2 d\omega < \infty. \tag{36}$$

Let $V_T^{(m)}$, $W_T^{(m)}$, $\Delta_{T,N,1}$, $\Delta_{T,N,2}$ and $\hat{N}$ be defined as in (26), (27), Lemma 2, and (35), respectively. Then we have

$$\left| W_T^{(m)} - \left(\begin{pmatrix} \Delta_1 & 0 \\ 0 & \Delta_2 \end{pmatrix} + \begin{pmatrix} 0 & \Gamma_{1,2} \\ \Gamma_{1,2}' & \Gamma_2 \end{pmatrix} \right) \right| \leq K \left\{ \frac{1}{\hat{N}} + \frac{1}{m} + \frac{1}{T_m} + \frac{m}{\hat{N}} \right\} \tag{37}$$

$$\left| V_T^{(m)} - \begin{pmatrix} \Delta_1 & 0 \\ 0 & \Delta_2 \end{pmatrix} \right| \leq K \frac{m}{\hat{N}}, \tag{38}$$

where K is a finite constant,

$$\begin{aligned} \Gamma_2 &= \frac{N}{T} \int_0^{T/N} \int_0^{2\pi} \int_0^{2\pi} \frac{\nabla_{\boldsymbol{\theta}} \mathcal{F}(v, \boldsymbol{\theta}_0, \omega_1) \nabla_{\boldsymbol{\theta}} \mathcal{F}(v, \boldsymbol{\theta}_0, \omega_1)'}{\mathcal{F}(v, \boldsymbol{\theta}_0, \omega_1)^2 \mathcal{F}(v, \boldsymbol{\theta}_0, \omega_2)^2} \\ &\quad \times \mathcal{F}_4(v, \boldsymbol{\vartheta}_0, \omega_1, \omega_2, -\omega_1) d\omega_1 d\omega_2 dv, \\ \Gamma_{1,2} &= \frac{N}{T} \int_0^{T/N} \mathbf{x}(v) \int_0^{2\pi} \int_0^{2\pi} \frac{\nabla_{\boldsymbol{\theta}} \mathcal{F}(v, \boldsymbol{\theta}_0, \omega_2)'}{\mathcal{F}(v, \boldsymbol{\theta}_0, \omega_1) \mathcal{F}(v, \boldsymbol{\theta}_0, \omega_2)^2} \\ &\quad \times \mathcal{F}_3(v, \boldsymbol{\vartheta}_0, \omega_1, \omega_2) \exp(ir\omega_2) \omega_1 d\omega_2 dv, \end{aligned} \tag{39}$$

$\mathbf{x}(v)' = (x_1(v), \ldots, x_n(v))$, with $\mathcal{F}(v, \boldsymbol{\theta}, \omega)$ defined as in (34),

$$\begin{aligned} \mathcal{F}_3(v, \boldsymbol{\vartheta}, \omega_1, \omega_2) &= \sum_{j=1}^{n+1} \kappa_{j,3} x_j(v)^3 A_j(\boldsymbol{\vartheta}, \omega_1) A_j(\boldsymbol{\vartheta}, \omega_2) A_j(\boldsymbol{\vartheta}, -\omega_1 - \omega_2) \\ \mathcal{F}_4(v, \boldsymbol{\vartheta}, \omega_1, \omega_2, \omega_3) &= \sum_{j=1}^{n+1} \kappa_{j,4} x_j(v)^4 A_j(\boldsymbol{\vartheta}, \omega_1) A_j(\boldsymbol{\vartheta}, \omega_2) A_j(\boldsymbol{\vartheta}, \omega_3) \\ &\quad \times A_j(\boldsymbol{\vartheta}, -\omega_1 - \omega_2 - \omega_3), \end{aligned}$$

$\kappa_{j,3} = cum(\eta_{0,j}, \eta_{0,j}, \eta_{0,j})$ and $\kappa_{j,4} = cum(\eta_{0,j}, \eta_{0,j}, \eta_{0,j}, \eta_{0,j})$.

REMARK 3 (Selecting m). Let us consider the case that $\{\alpha_{t,j}\}$ and $\{\varepsilon_t\}$ are Gaussian, in this case $\Gamma_{1,2} = 0$ and $\Gamma_2 = 0$. Comparing the asymptotic variances of the GMLE and $(\hat{\boldsymbol{a}}_T, \hat{\boldsymbol{\theta}}_T)$ we see that if we let $\hat{N} \to \infty$, $m \to \infty$, and $m/\hat{N} \to 0$ (noting that we have replaced $\hat{N}$ with N) as $T \to \infty$, then the GMLE $((\boldsymbol{a}_{mle}, \boldsymbol{\theta}_{mle}))$ and $(\hat{\boldsymbol{a}}_T, \hat{\boldsymbol{\theta}}_T)$ both have the same asymptotic distribution. Hence within this framework, the relative efficiency of the frequency domain estimator compared with the GMLE is one.

Furthermore, in the case that $\{\alpha_{t,j}\}$ and $\{\varepsilon_t\}$ are Gaussian, (37) suggests a method for selecting m. Since in this case the GMLE is efficient, by using (37) we have

$$\left| (V_T^{(m)})^{-1} W_T^{(m)} (V_T^{(m)})^{-1} - \operatorname{diag}(\Delta_1^{-1}, \Delta_2^{-1}) \right| = O_p\left(\frac{1}{\hat{N}} + \frac{1}{m} + \frac{1}{T_m} + \frac{m}{\hat{N}} \right).$$

Hence the above difference is minimized when $m = \hat{N}^{1/2}$.

6. Real data analysis

We now consider two real data examples.

Example 1. Application to financial time series: modeling of T-bills and inflation rates in the US

There are many possible applications of stochastic coefficient regression models in econometrics. One such application is modeling the influence of the nominal interest rate of 3-month (short term) Treasury bills (T-bills) on monthly inflation. Fama (1977) argues that the relationship between the T-bills and inflation rate determines whether the market for short-term Treasurey bills is efficient or not. In this section, we will consider 3-month T-bills and monthly inflation data observed monthly between January 1959 to December 2008, the data can be obtained from the US Federal reserve, http//www.federalreserve.gov/releases/h15/data.htm#fn26 and http://inflationdata.com/inflation/Inflation_Rate/HistoricalInflation.aspx, respectively. A plot of the time series of both sets of observations is given in Fig. 1. The estimated correlation coefficient between the 3-month T-bills and monthly inflation is 0.72. Let Y_t and x_t denote monthly inflation and T-bills interest rate at time t, respectively. Fama (1977) and Newbold and Bos (1985) consider the nominal interest rate of 3-month T-bills and inflation rate data observed every 3 months between 1953–1980. Fama (1977) fitted the linear regression model $Y_t = a_1 x_t + \varepsilon_t$ ($\{\varepsilon_t\}$ are i.i.d.) to the data, and showed that there was not a significant departure of a_1 from one,

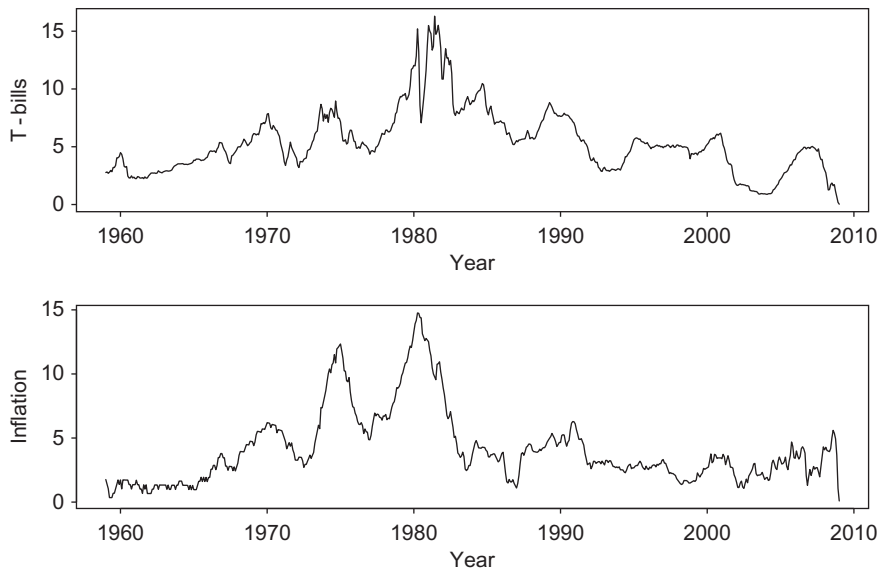

Fig. 1. The top plot is 3-month T-bill nominal interest rate taken monthly and lower plot is the monthly inflation rate.

he used this to argue that the T-bills market was efficient. However, Newbold and Bos (1985) argue that the relationship between T-bills and inflation is more complex and suggest that the SCR may be a more appropriate model, where the coefficient of x_t is stochastic and follows an AR(1) model. In other words

$$Y_t = a_0 + (a_1 + \alpha_{t,1})x_t + \varepsilon_t, \quad \alpha_{t,1} = \vartheta_1 \alpha_{t-1,1} + \eta_t \tag{40}$$

where $\{\varepsilon_t\}$ and $\{\eta_t\}$ are i.i.d. random variables with $\mathbb{E}(\varepsilon_t) = 0$, $\mathbb{E}(\eta_t) = 0$, $\text{var}(\varepsilon_t) = \sigma_\varepsilon^2 < \infty$, and $\text{var}(\eta_t) = \sigma_\eta^2 < \infty$. Using the GMLE they obtain the parameter estimates $a_0 = -0.97$, $a_1 = 1.09$, $\vartheta_1 = 0.89$, $\sigma_\varepsilon^2 = 1.41$, and $\sigma_\eta^2 = 0.013$. We now fit the same model to the T-bills data observed monthly from January 1959 to December 2008 (600 observations), and use the two-step estimator 2 to estimate the parameters a_0, a_1, ϑ_1, $\sigma_\alpha^2 = \text{var}(\alpha_{t,1})$, and $\sigma_\varepsilon^2 = \text{var}(\varepsilon_t)$ using estimator 2. The variances σ_α^2 and σ_ε^2 are estimated in the first step of estimator 2, the estimates with their standard errors are given in Table 1. Note that comparing the estimates with their standard errors, we observe that both parameters σ_ε^2 and σ_α^2 appear significant. In the second stage of the scheme, we estimate a_0, a_1, and ϑ_1 (we note that because the intercept a_0 appears to be insignificant and we also do the estimation excluding the intercept), these estimates are also summarized in Table 1. The estimates for different m are quite close. The model found

Table 1
We fit the model $Y_t = a_0 + (a_1 + \alpha_{t,1})x_t + \varepsilon_t$, where $\alpha_{t,1} = \vartheta_1 \alpha_{t-1,1} + \eta_t$, with and without the intercept a_0

	a_0	a_1	ϑ_1	σ_α	σ_ε
OLS	0.088	0.750			
(s.e.)	(0.18)	(0.029)			
Stage 1	0.088	0.74		0.285	1.083
(s.e.)				(0.011)	(0.059)
$m = 10$ (with intercept)	0.618	0.625	0.981		
(s.e.)	(0.325)	(0.069)	(0.042)		
$m = 10$ (without intercept)		0.741	(0.971)		
(s.e.)		(0.0325)	(0.05)		
$m = 50$ (with intercept)	0.309	0.687	0.969		
(s.e.)	(0.35)	(0.069)	(0.026)		
$m = 50$ (without intercept)		0.743	0.957		
(s.e.)		(0.032)	(0.038)		
$m = 200$ (with intercept)	0.223	0.7327	0.96088		
(s.e.)	(0.44)	(0.022)	(0.024)		
$m = 200$ (without intercept)		0.765	0.951		
(s.e.)		(0.029)	(0.030)		
$m = 400$ (with intercept)	0.367	0.725	0.963		
(s.e.)	(0.48)	(0.070)	(0.023)		
$m = 400$ (without intercept)		0.773	0.957		
(s.e.)		(0.029)	(0.026)		

The estimates using least squares and the frequency domain estimator for different m are given. The values in the brackets are the corresponding standard errors.

to be most suitable for the above data when $m = 200$ is

$$Y_t = (0.73 + \alpha_{t,1})x_t + \varepsilon_t, \quad \alpha_{t,1} = 0.960\alpha_{t-1,1} + \eta_t,$$

where $\sigma_\varepsilon^2 = 1.083^2$ and $\sigma_\alpha^2 = 0.285^2$ (hence $\sigma_\eta^2 = 0.079^2$). We observe that the AR(1) parameter estimate of the stochastic coefficient $\{\alpha_{t,1}\}$ is 0.96. This value is close to one, suggesting that the stochastic coefficients $\{\alpha_{t,1}\}$ could come from a unit root process.

To assess the validity of this model, we obtain one-step ahead best linear predictors of Y_t, given $\{Y_s\}_{s=1}^{t-1}$ and the current T-bills rate x_t, every month in the year 2008. In order to do the prediction we re-estimate the parameters α_1, θ, σ_ε^2, and σ_α^2 using the observations from 1959 to 2007. We use the two-step estimator 2 with $m = 200$ to obtain

$$Y_t = (0.77 + \alpha_{t,1})x_t + \varepsilon_t, \quad \alpha_{t,1} = 0.965\alpha_{t-1,1} + \eta_t, \tag{41}$$

with $\sigma_\varepsilon^2 = 0.79^2$ and $\sigma_\alpha^2 = 0.30^2$ (hence $\sigma_\eta^2 = 0.18^2$). We also fit the linear regression model $Y_t = a_1 x_t + \varepsilon_t$ to the data and use OLS to obtain the model $Y_t = 0.088 + 0.75x_t + \varepsilon_t$. The predictor using the usual multiple linear regression model and one-step ahead predictor using the SCR model are given in Fig. 2. To do the one-step ahead prediction we use the Kalman filter (using the R package `ss1.R` (see Shumway and Stoffer (2006), Chapter 6 for the details). The mean squared prediction errors over the 12 months using the multiple regression and the SCR model are 8.99 and 0.89, respectively. We observe from the plots in Fig. 2 that the multiple regression model always underestimates the true value and the mean square error is substantially larger than the SCR model.

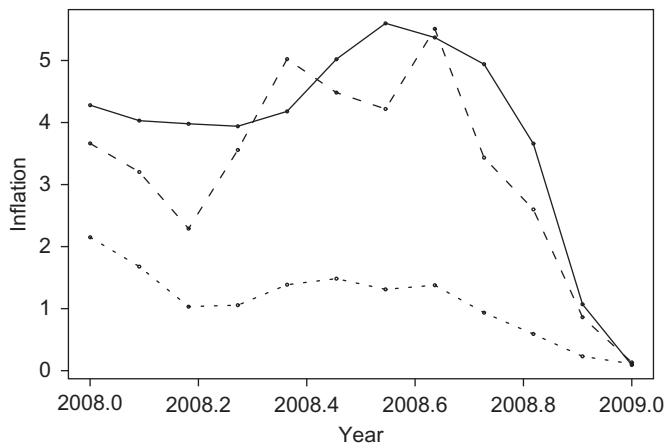

Fig. 2. We compare the true inflation rates with their predictions. The continuous thick line − is the true inflation rate. The broad dashed line −− is SCR one step ahead predictor given in (41). The fine dashed line ⋯ is the linear regression predictor $\hat{Y}_t = 0.088 + 0.750x_t$.

Example 2. Application to environmental time series: modeling of visibility and air pollution

It is known that air visibility quality depends on the amount of particulate matter (particulate matter negatively effects visibility). Furthermore, air pollution is known to influence the amount of particulate matter (see Hand and Malm (2008)). To model the influence of air pollution on particulate matter, Hand and Malm (2008) (see Eq. (6)) fit a linear regression model. However, Burnett and Guthrie (1970) argue that the influence of air pollution on air visibility may vary each day, depending on meteorological conditions, and suggest that a SCR model may be more appropriate than a multiple linear regression model. In this section we investigate this possibility. We consider the influence of man-made emissions on particulate matter (PM2.5-10) in Shenandoah National Park, Virginia, USA. The data we consider is ammonium nitrate extinction (ammNO3f), ammonium sulfate (ammSO4f), carbon elemental total (ECF), and particulate matter (PM2.5-10) (ammNO3f, ammSO4f, and ECF are measured in $\mu g/m^3$), which has been collected every 3 days between 2000 and 2005 (600 observations). We obtain the data from the VIEWS Web site `http://vista.cira.colostate.edu/views/Web/Data/DataWizard.aspx`. We mention that the influence of man-made emissions on air visibility is of particular importance to the US national parks service (NPS), who collected and compiled this data. An explanation of the data and how air pollution influences visibility (light scattering) can be found in the study by Hand and Malm (2008).

The plots of both the air pollution and PM2.5-10 data is given in Figs 3 and 4, respectively.

There is a clear seasonal component in all the data sets as seen from their plots. Therefore, to prevent spurious correlation between the PM2.5-10 and air pollution, we detrended and deseasonalized the PM2.5-10 and emissions data. To identify the dominating harmonics, we used the maximum periodogram methods suggested by Quinn and Fernandes (1991) and Kavalieris and Hannan (1994). To the detrended and deseasonalized PM2.5-10 and air pollution data we fitted the following model

$$Y_t = (a_1 + \alpha_{t,1})x_{t,1} + (a_2 + \alpha_{t,2})x_{t,2} + (a_3 + \alpha_{t,3})x_{t,3} + \varepsilon_t,$$

where $\{x_{t,1}\}$, $\{x_{t,2}\}$, $\{x_{t,3}\}$ and $\{Y_t\}$ are the detrended and deseasonalized ammNO3f, ammSO4f, ECF and Particulate Matter (PM2.5-10), and $\{\alpha_{t,j}\}$ and ε_t satisfy

$$\alpha_{t,j} = \vartheta_j \alpha_{t-1,j} + \eta_{t,j}, \quad \text{for and } j = 1, 2, 3, \qquad \varepsilon_t = \vartheta_4 \varepsilon_{t-1} + \eta_{t,4},$$

$\{\eta_{t,i}\}$ i.i.d. random variables. Let $\sigma_\varepsilon^2 = \text{var}(\varepsilon_t)$, $\sigma_{\alpha,1}^2 = \text{var}(\alpha_{t,1})$, $\sigma_{\alpha,2}^2 = \text{var}(\alpha_{t,2})$, and $\sigma_{\alpha,3}^2 = \text{var}(\alpha_{t,3})$.

We used estimator 2 to estimate the parameters $a_1, a_2, a_3, \alpha_{t,1}, \alpha_{t,2}, \alpha_{t,3}, \alpha_{t,4}, \sigma_\varepsilon^2, \sigma_{\alpha,1}^2$, $\sigma_{\alpha,2}^2$, and $\sigma_{\alpha,3}^2$. As initial values in the minimization scheme, we gave the least squares estimates of $\boldsymbol{a}_0$ for the mean regression coefficients and 0.1 for all the other unknown parameters. In the first stage of the scheme we estimated a_1, a_2, a_3 and $\sigma_\varepsilon^2, \sigma_{\alpha,1}^2, \sigma_{\alpha,2}^2$ and $\sigma_{\alpha,3}^2$. We also fitted a more parsimonious model where some of the coefficients were kept fixed rather than stochastic. The results are summarized in Table 2 (step 1).

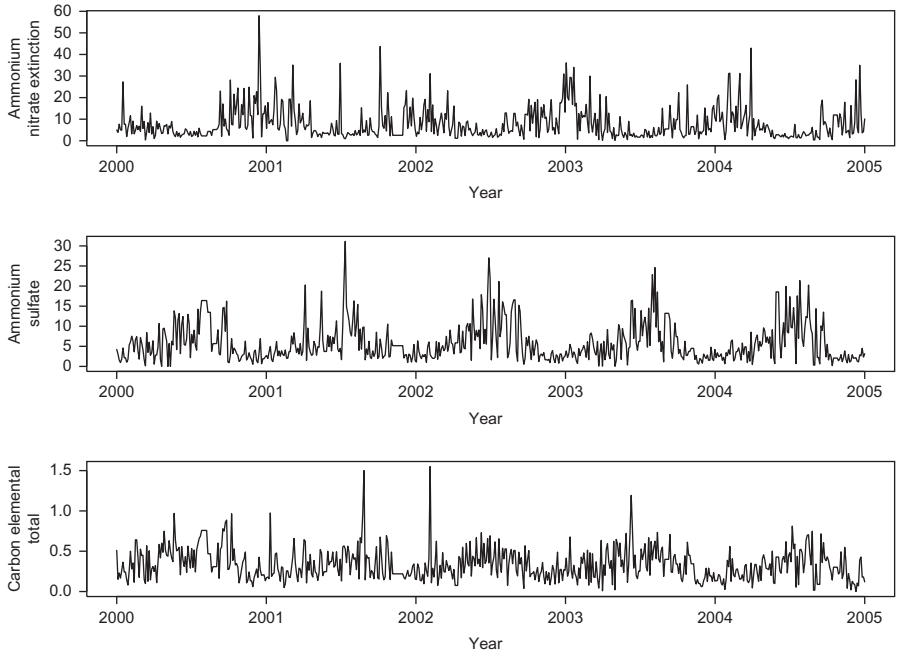

Fig. 3. The top plot is 3-day ammonium nitrate extinction (fine), middle plot is 3-day ammonium sulfate (fine), and lower plot is 3-day carbon elemental total (fine).

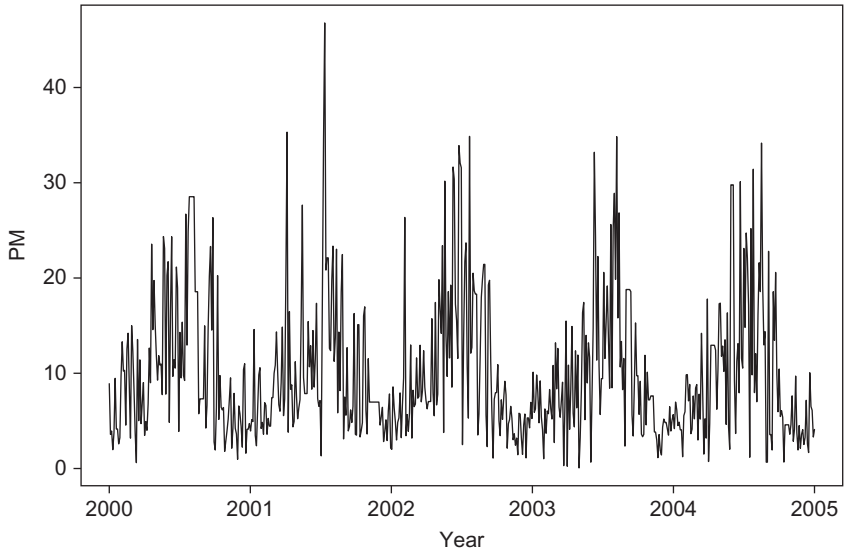

Fig. 4. The plot is 3-day particulate matter (PM2.5 - PM10).

Table 2
In stage 1 we fitted the model $Y_t = (a_1 + \alpha_{t,1})x_{t,1} + (a_2 + \alpha_{t,2})x_{t,2} + (a_3 + \alpha_{t,3})x_{t,3} + \varepsilon_t$, and various subsets (here we did not model any dependence in the stochastic coefficients)

	a_1	a_2	a_3	ϑ_2	ϑ_3	ϑ_4	$\sqrt{\text{var}(\alpha_{t,1})}$	$\sqrt{\text{var}(\alpha_{t,2})}$	$\sqrt{\text{var}(\alpha_{t,3})}$	$\sqrt{\text{var}(\varepsilon_t)}$	minL
OLS	0.29	4.57	1.76								
	(0.078)	(0.088)	(0.0908)								
Stage 1	0.38	4.53	1.58				7.10^{-7}	1.25	0.84	1.29	2.102
(s.e.)	(0.048)	(0.079)	(0.076)				(0.07)	(0.097)	(0.098)	(0.046)	
Stage 1	0.56	3.28	−3.09				0.75		3.315	4.496	4.09
(s.e.)	(0.170)	(0.181)	(0.29)				(0.042)		(0.159)	(0.049)	
Stage 1	0.387	4.53	1.58					1.157	0.84	1.296	2.102
(s.e.)	(0.048)	(0.080)	(0.076)					(0.009)	(0.009)	(0.002)	
Stage 1	0.287	3.52	−3.11						3.00	3.83	3.99
(s.e.)	(0.139)	(0.171)	(0.27)						(0.164)	(0.07)	
Stage 1	0.521	5.09	−2.93							4.42	4.15
(s.e.)	(0.131)	(0.145)	(0.65)							(0.011)	
$m = 10$	0.393	4.47	1.630	0.883	−0.144						2.066
(s.e.)	(0.048)	(0.079)	(0.076)	(0.095)	(0.351)						
$m = 10$	0.39	4.47	1.63		−0.13						2.066
(s.e.)	(0.05)	(0.077)	(0.071)		(0.34)						
$m = 10$	0.388	4.467	1.661			0.94					2.084
(s.e.)	(0.039)	(0.061)	(0.058)			(0.029)					

(Continued)

Table 2
(Continued)

	a_1	a_2	a_3	ϑ_2	ϑ_3	ϑ_4	$\sqrt{\text{var}(\alpha_{t,1})}$	$\sqrt{\text{var}(\alpha_{t,2})}$	$\sqrt{\text{var}(\alpha_{t,3})}$	$\sqrt{\text{var}(\varepsilon_t)}$	minL
$m = 50$	0.327	4.55	1.75	0.617	−0.235						2.199
(s.e.)	(0.056)	(0.070)	(0.071)	(0.30)	(0.659)						
$m = 50$	0.324	4.54	1.746		−0.32						2.22
(s.e.)	(0.055)	(0.073)	(0.070)		(0.54)						
$m = 50$	0.308	4.55	1.764			0.244					2.21
(s.e.)	(0.049)	(0.062)	(0.062)			(0.127)					
$m = 200$	0.261	4.595	1.770	0.538	0.9722						2.111
(s.e.)	(0.06)	(0.067)	(0.070)	(0.322)	(0.042)						
$m = 200$	0.261	4.60	1.77		−0.087						2.124
(s.e.)	(0.06)	(0.068)	(0.068)		(1.2)						
$m = 200$	0.255	4.58	1.797			0.458					2.1339
(s.e.)	(0.051)	(0.058)	(0.058)			(0.145)					
$m = 400$	0.2531	4.597	1.793	0.979	0.932						2.116
(s.e.)	(0.061)	(0.068)	(0.070)	(0.032)	(0.143)						
$m = 400$	0.25	4.59	1.79		0.92						2.128
(s.e.)	(0.06)	(0.068)	(0.068)		(0.167)						
$m = 400$	0.250	4.580	1.807			0.478					2.139
(s.e.)	(0.051)	(0.0583)	(0.059)			(0.148)					

In the second stage (for $m = 10, 50, 200$, and 400) we fitted AR(1) models to $\{\alpha_{t,2}\}$, $\{\alpha_{t,3}\}$, and $\{\varepsilon_t\}$, that is $\alpha_{t,2} = \vartheta_2\alpha_{t-1,2} + \eta_{t,2}$, $\alpha_{t,3} = \vartheta_3\alpha_{t-1,3} + \eta_{t,3}$ and $\varepsilon_t = \vartheta_4\varepsilon_{t-1} + \eta_{t,4}$. The value of the frequency domain likelihood at the minimal value is also given in the column minL. The standard errors of the estimates are given below each estimate in brackets.

We observe that the estimate of $\sigma_{\alpha,1}$ is extremely small and insignificant, and that the minimum value of the objective function $\mathcal{L}_T$ is about the same when the coefficient of ammNO3f $\{x_{t,1}\}$ is fixed and random (it is 2.012). This suggests that the coefficient of ammNO3f $\{x_{t,1}\}$ is deterministic. This may indicate that the relative contribution of NO3f ($\{x_{t,1}\}$) to the response is constant throughout the period of time and is not influenced by any other extraneous factors. We re-did the minimization systematically removing $\sigma_{\alpha,2}$ and $\sigma_{\alpha,3}$, but the minimum value of $\mathcal{L}_T$, changed quite substantially (compare the minimum of the objective functions 2.012 with 4.09, 3.99 and 4.15). Hence the most appropriate model appears to be

$$Y_t = a_1 x_{t,1} + (a_2 + \alpha_{t,2})x_{t,2} + (a_3 + \alpha_{t,3})x_{t,3} + \varepsilon_t,$$

where $\{\alpha_{t,1}\}$ and $\{\alpha_{t,2}\}$ are stochastic coefficients. It is of interest to investigate whether the coefficients of ammSO4f and ECF are purely random or correlated, and we investigate this in the second stage of the frequency domain scheme, where we modeled $\{\alpha_{t,2}\}$ and $\{\alpha_{t,3}\}$ both as i.i.d. random variables and as the AR(1) model $\alpha_{t,j} = \vartheta_j \alpha_{t-1,j} + \eta_{t,j}$, for $j = 2, 3$. The estimates for various different models and different values of m are given in Table 2. If we compare the minimum of the objective function where $\{\alpha_{t,2}\}$ and $\{\alpha_{t,3}\}$ are modeled as both i.i.d. and satisfying an AR(1) model, we see that there is very little difference between them. Moreover, the standard errors for the estimates of ϑ_2 and ϑ_3 are large. Altogether, this suggests that ϑ_2 and ϑ_3 are not significant and $\{\alpha_{t,2}\}$ and $\{\alpha_{t,3}\}$ are uncorrelated over time. Hence, it seems plausible that the coefficients of ammSO4f and ECF are random, but independent. To check the possibility that the errors $\{\varepsilon_t\}$ are correlated, we fitted an AR(1) model to the errors. However we observe from Table 2, that the AR(1) parameter does not appear to be significant. Moreover, comparing the minimum of the objective function $\mathcal{L}_{600}^{(m)}$ (for different values of m) fitting i.i.d. $\{\varepsilon_t\}$ and an AR(1) to $\{\varepsilon_t\}$ gives almost the same value. This suggests that the errors are independent. In summary, our analysis suggests that the influence of ammNO3f on PM2.5-10 is fixed over time, whereas the influence of ammSO4f and ECF varies purely randomly over time. Using the estimator obtained when $m = 200$ this suggests the model

$$Y_t = 0.255 x_{t,1} + (4.58 + \alpha_{t,2})x_{t,2} + (1.79 + \alpha_{t,3})x_{t,3} + \varepsilon_t,$$

where $\{\alpha_{t,2}\}$ and $\{\alpha_{t,3}\}$ are i.i.d. random variables, with $\sigma_{\alpha,2} = 1.157$, $\sigma_{\alpha,3} = 0.84$, and $\sigma_\varepsilon = 1.296$. Based on our analysis it would appear that the coefficients of pollutants are random, but there is no linear dependence between the current coefficient and the previous coefficient. On possible explanation for the lack of dependence is that the data is taken every 3 days and not daily. This could mean that the meteorological conditions from 3 days ago has little influence on today's particulate matter. On the other hand if we were to analyze the daily pollutants and daily PM2.5-10 the conclusions could have been different. But this daily data is not available. It is likely that since the data is aggregated (smoothed) over a 3-day period any possible dependence in the data was removed.

Acknowledgments

The author wishes to thank Dr. Bret Schichtel, at CIRA, Colorado State University for his invaluable advice and explanations of the VIEWs data. This work has been partially supported by the NSF Grant DMS-0806096 and the Deutsche Forschungsgemeinschaft DA 187/15-1.

References

Breusch, T.S., Pagan, A.R., 1980. A simple test for for heteroscedasticity and random coefficient variation. Econmetrica 47, 1287–1294.

Burnett, T.D., Guthrie, D., 1970. Estimation of stationary stochastic regression parameters. J. Am. Stat. Assoc. 65, 1547–1553.

Caines, P., 1988. Linear Stochastic Systems. Wiley, New York.

Dahlhaus, R., 1996. Maximum likelihood estimation and model selection for locally stationary processes. J. Nonparametri. Stat. 6, 171–191.

Dahlhaus, R., 2000. A likelihood approximation for locally stationary processes. Ann. Stat. 28, 1762–1794.

Duncan, D.B., Horn, S., 1972. Linear dynamic recursion estimation from the viewpoint of regression analysis. J. Am. Stat. Assoc. 67, 815–821.

Dunsmuir, W., 1979. A central limit theorem for parameter estimation in stationary vector time series and its application to models for a signal observed with noise. Ann. Stat. 7, 490–506.

Dwivedi, Y., Subba Rao, S., 2011. A test for second order stationarity based on the discrete fourier transform. J. Time Ser. Anal. 32, 68–91.

Dzhapharidze, K., 1971. On methods for obtaining asymptotically efficient spectral parameter estimates for a stationary Gaussian processes with rational spectral density. Theory Probab. Appl. 16, 550–554.

Fama, E.F., 1977. Interest rates and inflation: the message in the entrails. Ame. Econ. Rev. 67, 487–496.

Franke, J., Gründer, B., 1995. General kriging for spatial-temporal processes with random ARX-regression parameters. In: Robinson, P.M., Rosenblatt, M. (Eds.), Athens Conference in Applied Probability and Time Series Analysis, vol. ii. Springer, New York, pp. 177–189.

Giraitis, L., Robinson, P., 2001. Whittle estimation of ARCH models. Econom. Theory 17, 608–631.

Hand, J.L., Malm, W.C., 2008. Review of the improve equation for estimating ambient light extinction coefficients - final report (Tech. Rep.). http://vista.cira.colostate.edu.

Hannan, E.J., 1971. Non-linear time series regression. J. Appl. Probab. 8, 767–780.

Hannan, E.J., 1973. The asymptotic theory of linear time series models. J. Appl. Probab. 10, 130–145.

Hildreth, C., Houck, C., 1968. Some estimates for a linear model with random coefficients. J. Am. Stat. Assoc. 63, 584–595.

Kavalieris, L., Hannan, E.J., 1994. Determining the number of terms in a periodic regression. J. Time Ser. Anal. 15, 6130625.

Ljung, L., Caines, P., 1979. Asymptotic normality of prediction error estimators for approximate system models. Stochastics 3, 29–46.

Martinussen, T., Scheike, T.H., 2000. A nonparametric dynamic additive regression model for longitudinal data. Ann. Stat. 28, 1000–1025.

Newbold, P., Bos, T., 1985. Stochastic Parameter Regression Models. Sage Publications, Beverly Hills.

Pfeffermann, D., 1984. On extensions of the Gauss-Markov theorem to the case of stochastic-regression coefficient. J. R. Stat. Soc. B 46, 139–148.

Picinbono, B., 1996. Second-order complex random vectors and normal distributions. IEEE Trans. Signal Process. 44, 2637–2640.

Quinn, B.G., Fernandes, J.M., 1991. A fast and efficient technique for the estimation of frequency. Biometrika 78, 489–498.

Rosenberg, B., 1972. The estimation of stationary stochastic regression parameters reexamined. J. Am. Stat. Assoc. 67, 650–654.

Rosenberg, B., 1973. Linear regression with randomly dispersed parameters reexamined. Biometrika 60, 65–72.

Shumway, R.H., Stoffer, D.S., 2006. Time Series Analysis and Its Applications (with R examples). Springer, New York.
Stoffer, D.S., Wall, K.D., 1991. Bootstrapping state space models: Gaussian maximum likelihood estimation. J. Am. Stat. Assoc. 86, 1024–1033.
Swamy, P.A.V.B., 1970. Efficient inference in a random coefficient regression model. Econometrica 38, 311–323.
Swamy, P.A.V.B., 1971. Statistical Inference in Random Coefficient Models. Springer, Berlin.
Swamy, P.A.V.B., Tinsley, P.A., 1980. Linear prediction and estimation methods for regression model with stationary stochastic coefficients. J. Econom. 12, 103–142.
Synder, R.D., 1985. Recursive estimation of dynamic linear models. J. R. Stat. Soc. B 47, 272–276.
Taniguchi, M., 1983. On the second order asymptotic efficiency of estimators of Gaussian ARMA processes. Ann. Stat. 11, 157–169.
Walker, A.M., 1964. Asymptotic properties of least squares estimates of parameters of the spectrum of nonstationary non-deterministic time series. J. Aust. Math. Soc. 4, 363–384.
Whittle, P., 1962. Gaussian estimation in stationary time series. Bull. Int. Stat. Inst. 39, 105–129.

Part VII: Spatio-Temporal Time Series

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00016-8

Hierarchical Bayesian Models for Space–Time Air Pollution Data

Sujit K. Sahu

School of Mathematics, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK

Abstract

Recent successful developments in Bayesian modeling and computation for large and complex data sets have led to a step change in the analysis of space–time air pollution data. Accurate predictions and inferences, as a result of joint modeling of spatial and temporal dependence, are being made even for summaries and aggregates in time and/or space. Modelers are increasingly benefiting from their ability to reducing uncertainty in the inference statements for the aggregates, in addition, by combining information from several sources in a hierarchical Bayesian framework. The information sources may include actual observed data from several heterogeneous monitoring networks, outputs of community numerical models, meteorological observations, land use surfaces, and power station emission volumes. The best statistical model for the particular problem at hand recognizes the relative contribution of the available sources and decides on their optimum role in it. In this chapter, we develop a hierarchical autoregressive Bayesian model for space–time air pollution data and illustrate the benefits of modeling with a real data example on monitoring ozone pollution. We report substantial gains in predictive mean square error for the proposed model over some other currently available competing modeling methods.

Keywords: Auto-regressive models, Criteria pollutants, Monitoring compliance, Ozone concentration modeling, Spatial interpolation.

1. Introduction

The clean air act, amended in 1990 by the legislators in the United States, requires two types of air quality standards to be maintained for six most important air pollutants: ozone, particulate matter, carbon monoxide, nitrogen oxides, sulfur dioxide,

and lead. The first, *primary* standards, set limits to protect public health, including the health of "sensitive" populations such as asthmatics, children, and the elderly. The *secondary* standards set limits to public welfare, provide protection against decreased visibility, damage to animals, crops, vegetation, and buildings. The Website *http://epa.gov/air/criteria.html* provides the current values of these standards. Out of the six pollutants, particulate matter and ozone have received the most attention in the literature and these two are the focus of this chapter, although the modeling methods are applicable to other pollutants as well.

To monitor compliance to the air quality standards and to evaluate exposure to air pollution the United States Environmental Protection Agency (USEPA) has developed several sparse networks of monitoring stations covering the whole of the United States. Data obtained from these sparse networks must be processed using stochastic spatial and spatio-temporal models to make valid inference for air pollution levels at particular sites, such as urban areas, based on rigorous statistical methods. In fact, the demand for spatial models to assess regional progress in air quality has grown rapidly over the past decade. For improved environmental decision making, it is imperative that such models enable spatial prediction to reveal important gradients in air pollution, offer guidance for determining areas in nonattainment with air quality standards, and provide air quality input to models for determining individual exposure to air pollution. Spatial prediction has the potential to suggest new perspectives in the development of emission control strategies and to provide a credible basis for resource allocation decisions, particularly with regard to network design.

Particulate matter is a complex mixture of extremely small particles and liquid droplets, and is harmful to human health when inhaled. There are two variations of particulate matter. Particles with diameters less than 10 micrometers (μm) are called PM_{10} and those with diameters less than 2.5 μm are called $PM_{2.5}$. PM_{10} are found near roadways and dusty industries, and $PM_{2.5}$ are found in smoke and haze. A lot of modeling effort has gone into analyzing spatio-temporal behavior of particulate matter, below we provide a brief review. Cressie et al. (1999) compare kriging and Markov-random field models in the prediction of PM_{10} concentrations around the city of Pittsburgh. Sun et al. (2000) develop a spatial predictive distribution for the space–time response of daily ambient PM_{10} in Vancouver, Canada. Kibria et al. (2002) develop a multivariate spatial prediction methodology in a Bayesian context for the prediction of $PM_{2.5}$ in the city of Philadelphia. This approach used both $PM_{2.5}$ and PM_{10} data at monitoring sites with different start-up times. Shaddick and Wakefield (2002) propose short term space–time modeling for PM_{10}. Zidek et al. (2002) develop predictive distributions on nonmonitored PM_{10} concentrations in Vancouver. Smith et al. (2003) propose a spatio-temporal model for predicting weekly averages of $PM_{2.5}$ and other derived quantities such as annual averages within three southeastern states in the US. Sahu and Mardia (2005) present a short-term forecasting analysis of $PM_{2.5}$ data in New York City during 2002. Sahu et al. (2006) consider modeling of $PM_{2.5}$ by mixing two processes: one for the rural background areas and the other for the urban areas. Cocchi et al. (2007) develop hierarchical Bayesian model for daily average PM_{10} concentration levels. Pollice and Lasinio (2010) develop a Bayesian kriging-based method for estimating daily PM_{10} surfaces.

Ground-level ozone is a pollutant that is a significant health risk, especially for children with asthma and vulnerable adults with respiratory problems. It also damages

crops, trees, and other vegetation. It is a main ingredient of urban smog. Some early references on ozone modeling include Cox and Chu (1992), Brown et al. (1994), Guttorp et al. (1994), Carroll et al. (1997), and Thompson et al. (2001). Porter et al. (2001) report on the estimation of trends in ozone concentrations adjusted for meteorological variables at individual monitoring sites. Zhu et al. (2003) relate ambient ozone and pediatric asthma emergency room visits in Atlanta using hierarchical regression methods for spatially misaligned data. Huerta et al. (2004) model hourly readings of concentrations of ozone jointly with air temperature for data from Mexico City. Cocchi et al. (2005) follow the approach of Huang and Smith (1999) by using a tree-based partitioning of daily maxima ozone concentrations and assumed these maxima are Weibull distributed. McMillan et al. (2005) propose a regime-switching model for ozone forecasting using meteorological variables as covariates and they illustrate using data from April to September in 1999 over a spatial domain covering Lake Michigan. Sahu et al. (2007) deal with misalignment between ozone data and meteorological information. Sahu et al. (2009b) develop a hierarchical space–time model for daily 8-h maximum ozone concentration data covering much of the eastern United States. Dou et al. (2010) compare two Bayesian methods for modeling hourly ozone concentration levels. Berrocal et al. (2010, 2011) propose various downscaling approaches by regressing the observed point level ozone concentration data on grid cell level computer model output with spatially varying regression coefficients specified through a Gaussian process.

Several authors have developed generic models for analyzing spatio-temporal data. Research in this area dates back to Cressie (1994), Goodall and Mardia (1994), and Mardia et al. (1998). More recent articles in this area include: Kyriakidis and Journal (1999), Stroud et al. (2001), Wikle and Cressie (1999), Wikle (2003), Gelfand et al. (2005), and Cressie et al. (2010). A recent book, Cressie and Wikle (2011), provides a very comprehensive review of both classical and Bayesian methods for analysing space–time data.

The format of the remainder of this chapter is as follows. In Section 2, we develop the hierarchical autoregressive model based on our recent work. We also provide an introduction to Gaussian processes. Spatial prediction and forecasting methods, including their estimation in an iterative Markov chain Monte Carlo (MCMC) computation setup, are provided in Section 3. An illustration of the modeling methods is given in Section 4 using daily maximum 8-hour average ozone concentration levels observed in three mid-western states namely, Illinois, Indiana, and Ohio in 2006. A few summary remarks are provided in Section 5. The Appendix outlines the full conditional distributions needed for setting up the MCMC.

2. Hierarchical models

2.1. Models for data

Point level air pollution data at location $\mathbf{s}$ and at time t is denoted by $Z(\mathbf{s}, t)$, after any transformation, if necessary. Air pollution data are often modeled on the square-root scale, which encourages normality and stabilizes the variance, see, e.g., Sahu et al. (2007), although the log transformation is also used sometimes. Model fitting statistics,

i.e., goodness-of-fit, model diagnostics, parameter estimates and their uncertainty measures are reported on the modeled scale. However, the validations and predictions are reported on the original scale for ease of communication to the practitioners and the end users.

We assume that $Z(\mathbf{s},t)$ is univariate and the spatial reference vector $\mathbf{s}$ is two-dimensional describing the latitude–longitude pair (or its equivalent Northing and Easting coordinates for example) and the time index t is discrete. We also assume that $Z(\mathbf{s},t)$ is observed at n monitoring sites denoted by $\mathbf{s}_i$, $i=1,\ldots,n$, say and at T time points so that $t=1,\ldots,T$. The time unit is typically an hour or a day, although coarser units such as month or year are also used depending on the specific modeling objectives.

The first stage of the hierarchy assumes the measurement error model:

$$Z(\mathbf{s}_i,t)=Y(\mathbf{s}_i,t)+\epsilon(\mathbf{s}_i,t),\ i=1,\ldots,n,\quad t=1,\ldots,T, \tag{1}$$

where $Y(\mathbf{s}_i,t)$ is the true underline spatio-temporal process and the error term $\epsilon(\mathbf{s}_i,t)$ is a white noise process, assumed to follow the $N(0,\sigma_\epsilon^2)$ distribution. In the spatial statistics literature, σ_ϵ^2 is often called the nugget effect that quantifies variation of the data points measured at locations that are very small distances apart. In principle, σ_ϵ^2 could evolve in time but in many applications it is treated as a constant for the sake of parsimony. This first stage specification is advantageous in handling missing data in the Bayesian modeling setup with MCMC computation, since any missing data $Z(\mathbf{s}_i,t)$ is simply simulated from the $N(Y(\mathbf{s}_i,t),\sigma_\epsilon^2)$ distribution as implied by (1) at each MCMC iteration. The specification for $Y(\mathbf{s}_i,t)$ is provided in the next stage.

The space–time process $Y(\mathbf{s}_i,t)$ is assumed to have a systematic mean component, $\mu(\mathbf{s}_i,t)$ that may depend on past values and relevant covariates. A first-order autoregressive model given by $\rho Y(\mathbf{s}_i,t-1)$ can be used to model dependence on past values. This model will be appropriate when there is high autocorrelation present between the successive temporal values at any particular site. Additional autoregressive terms can also be considered if the first order model is inadequate in modeling the temporal dependence. Those terms, however, may not remain significant when other model components such as the covariate effects are introduced. In this chapter, we will only consider the first-order autoregressive process for the sake of parsimony.

The mean function, $\mu(\mathbf{s}_i,t)$, can be further enriched by a set of p, say, relevant spatially and temporally varying covariates $\mathbf{x}(\mathbf{s}_i,t)=(x_1(s_i,t),\ldots,x_p(s_i,t))'$. Note that some of these can only vary spatially and some others may only vary temporally. The covariate effect can be assumed to be spatially varying by assuming a spatially varying p-dimensional coefficient process $\boldsymbol{\beta}(\mathbf{s})$. This model allows the possibility of particular covariates making local adjustments to the mean function. A suitable prior process must be specified for $\boldsymbol{\beta}(\mathbf{s})$. Below we discuss a Gaussian process prior often used in practical problems. In our illustration in Section 4, however, we will use a fixed $\boldsymbol{\beta}$ for all sites $\mathbf{s}$.

The third and final component in $Y(\mathbf{s}_i,t)$ is assumed to be a residual random intercept, $w(\mathbf{s}_i,t)$, varying in both space and time. Having modeled the temporal dependence by an autoregressive process, we can assume $w(\mathbf{s}_i,t)$ to be a temporally independent zero-mean Gaussian process with a specified covariance function. This independence assumption simplifies the computation since covariance matrices of only order $n\times n$ need to be worked with instead of the full $nT\times nT$ matrices. Wikle and

Cressie (1999) suggest an alternative specification for $w(\mathbf{s}_i, t)$ using orthonormal basis functions in space and random mean zero variables in time.

In summary, the second-stage model specification is given by:

$$Y(\mathbf{s}_i, t) = \rho Y(\mathbf{s}_i, t-1) + \mathbf{x}(\mathbf{s}_i, t)'\boldsymbol{\beta} + w(\mathbf{s}_i, t). \tag{2}$$

Note that the autoregressive component in time and the regression term compete against each other to provide alternative explanations of data. These together also compete against the explanation provided by the assumed spatial correlation structure. Because of this, a practical model fitting exercise can be thought to weigh-up these three alternative sources of information for choosing the best possible mixture of model components for explaining the data. Of course, formal Bayesian model choice criteria can be adopted to compare specific models of interest, i.e., the model without the regressors and so on. For models based on a top-level Gaussian distribution there are several predictive Bayesian model choice criteria such as the predictive model choice criteria (PMCC), see, e.g., Sahu et al. (2009b) for an illustration with the PMCC. In this chapter, however, we do not consider such model choice criteria any further and instead use the significance of parameter estimates to decide whether to include them in the model.

The autoregressive model requires a specification for the initial condition $\mathbf{Y}_0' = (Y(\mathbf{s}_1, 0), \ldots, Y(\mathbf{s}_n, 0))$. There are two possible alternative specifications for $\mathbf{Y}_0$, e.g., (i) treat it as a fixed constant where $Y(\mathbf{s}_i, 0)$ is set at the overall mean of location $\mathbf{s}_i$, (ii) assign a prior distribution with mean $\boldsymbol{\mu}_0$ and covariance matrix Σ_0. In the latter case, the elements of $\boldsymbol{\mu}_0$ can be taken as the sitewise means, but there may be several possibilities for treating Σ_0. For example, it may be assumed to be a diagonal matrix with a large value 10^4, say for each diagonal entry corresponding to the assumption of a flat prior. Alternatively, elements of Σ_0 can be specified using a Gaussian covariance function discussed in the next subsection. In our illustration, we treat $\mathbf{Y}_0$ as fixed for convenience and simplicity.

2.2. *Gaussian processes*

Often Gaussian processes are assumed as components in spatial and spatio-temporal modeling. These stochastic processes are defined over a continuum, e.g., a spatial study region and specifying the resulting infinite dimensional random variable is often a challenge in practice. Gaussian processes are very convenient to work in these settings since they are fully defined by a mean function, say $\mu(\mathbf{s})$ and a valid covariance function, say $C(s, s^*) = \mathrm{Cov}(w(s), w(s^*))$, which is required to be positive definite. A covariance function is said to be positive definite if the covariance matrix, implied by that covariance function, for a finite number of random variables belonging to that process is positive definite. Below we provide a family of valid positive definite covariance functions.

Gaussian processes are often preferred in spatial modeling because of the attractive distribution theory associated with them. All finite dimensional distributions of Gaussian processes are multivariate normals. Hence, joint distribution of data observed at any finite set of locations (or the associated random effects) is multivariate normal. Moreover, kriging or spatial prediction at yet unobserved locations conditionally on the observed data is facilitated by means of a conditional distribution, which is also normal.

This convenient distribution theory is very attractive for spatial prediction in the context of modern, fully model-based spatial analysis within a Bayesian framework. The spatial predictive distributions are easy to compute and simulate from in an iterative MCMC framework.

We now turn to the specification of a valid covariance function. There is a substantial literature on this. Chapter 2 of Banerjee et al. (2004) provides a thorough discussion on this topic and the related concepts of stationarity, isotropy, and separability. Briefly, a process is defined to be *weakly stationary* if the covariance between a pair of random variables depends only on the separation vector between the two locations and not on the actual locations where those are observed. An *isotropic* covariance function only depends upon the distance between any two locations and not on the direction. Thus any pair of random variables observed at any two locations will have the same covariance as any other pair of random variables observed at any other locations separated by same distance. Separability is a concept used in modeling multivariate spatial data including spatio-temporal data. A *separable covariance function* in space and time is simply the product of two covariance functions: one for space and the other for time.

The Matérn family of covariance functions provides a very general choice and is given by:

$$C(u) = \sigma^2 \frac{1}{2^{\nu-1}\Gamma(\nu)} (2\sqrt{\nu} u\phi)^{\nu} K_{\nu}(2\sqrt{\nu} u\phi), \quad \phi > 0, \nu \geq 1, u > 0, \tag{3}$$

where $K_{\nu}(\cdot)$ is the modified Bessel function of second kind and of order ν, see, e.g., Abramowitz and Stegun (1965, Chapter 9). Popular special cases of the Matérn family are: (i) $\nu = 1/2$ corresponding to the exponential model $C(u) = \sigma^2 \exp(-\phi u)$ and (ii) $\nu = 3/2$ which leads to $C(u) = \sigma^2(1 + \phi u)\exp(-\phi u)$ and (iii) Gaussian, $C(u) = \sigma^2 \exp(-\phi^2 u^2)$ when $\nu \to \infty$.

The minimum value of u for which $C(u) \approx 0$ is defined as the *range* in spatial statistics literature. Note that, for the exponential covariance function, $C(u)$ can be exactly 0 only when u is very large, in other words ∞. To avoid this value of infinite range when our study region is a finite domain (in the sense that the maximum distance between any two locations is finite), we often calculate the range as that value of the distance u for which $C(u)$ is very small, i.e., 0.01 or 0.05. In this chapter, we shall illustrate with the exponential covariance function for which we define the range as $-\log(0.05)/\phi \approx 3/\phi$.

2.3. *Joint posterior distribution*

Define $\mathbf{Z}_t = (Z(\mathbf{s}_1, t), \ldots, Z(\mathbf{s}_n, t))'$, and $\mathbf{Y}_t = (Y(\mathbf{s}_1, t), \ldots, Y(\mathbf{s}_n, t))'$. Let X_t denote the $n \times p$ matrix having the ith row as $\mathbf{x}(\mathbf{s}_i, t)'$. It is convenient to write the joint posterior distribution using $\mathbf{Z}_t$, $\mathbf{Y}_t$, and X_t. To facilitate this, we now rewrite the hierarchical model specifications using these vectors and matrices as follows. The first model equation is obtained from (1):

$$\mathbf{Z}_t = \mathbf{Y}_t + \boldsymbol{\epsilon}_t, \tag{4}$$

for $t = 1, \ldots, T$, where $\boldsymbol{\epsilon}_t = (\epsilon(\mathbf{s}_1, t), \ldots, \epsilon(\mathbf{s}_n, t))'$. From (2) we have:

$$\mathbf{Y}_t = \rho\, \mathbf{Y}_{t-1} + X_t \boldsymbol{\beta} + \mathbf{w}_t, \tag{5}$$

for $t = 1, \ldots, T$, where $\mathbf{w}_t = (w(\mathbf{s}_1, t), \ldots, w(\mathbf{s}_n, t))'$.

For the measurement error model in (4) we have that $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \sigma_\epsilon^2 I_n)$, $t = 1, \ldots, T$, independently, where $\mathbf{0}$ is the vector with all elements zero and I_n is the identity matrix of order n. For the spatially correlated error we assume that $\mathbf{w}_t$ follows the GP independently with the covariance function $\sigma_w^2 \rho_w(\mathbf{s}_i - \mathbf{s}_j; \phi_w)$. We take $\rho_w(\mathbf{s}_i - \mathbf{s}_j; \phi_w) = \exp\left(-\phi_w d(\mathbf{s}_i, \mathbf{s}_j)\right)$, where $d(\mathbf{s}_i, \mathbf{s}_j)$ is the distance between sites $\mathbf{s}_i$ and $\mathbf{s}_j$, $i, j = 1, \ldots, n$. This GP assumption implies that $\mathbf{w}_t \sim N(\mathbf{0}, \Sigma_w)$, $t = 1, \ldots, T$, where Σ_w has elements $\sigma_w(i, j) = \sigma_w^2 \exp\left(-\phi_w d(\mathbf{s}_i, \mathbf{s}_j)\right)$. For future use, we define S_w by the relation $\Sigma_w = \sigma_w^2 S_w$.

Let $\boldsymbol{\vartheta}_t = \rho\, \mathbf{Y}_{t-1} + X_t \boldsymbol{\beta}$, for $t = 1, \ldots, T$. Further, let $\boldsymbol{\theta}$ denote all the parameters, $\boldsymbol{\beta}$, ρ, σ_ϵ^2, ϕ_w, and σ_w^2. Let $\mathbf{v}$ denote all the augmented data, $\mathbf{Y}_t$ and the missing data, denoted by $z^*(\mathbf{s}_i, t)$, for $i = 1, \ldots, n$, $t = 1, \ldots, T$, and $\mathbf{z}$ denote all the observed non-missing data $z(\mathbf{s}_i, t)$, for $i = 1, \ldots, n$, $t = 1, \ldots, T$. The log of the posterior distribution, denoted by $\log \pi(\boldsymbol{\theta}, \mathbf{v}|\mathbf{z})$, can be written as

$$\begin{aligned}
&-\frac{nT}{2} \log(\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2} \sum_{t=1}^{T} (\mathbf{Z}_t - \mathbf{Y}_t)'(\mathbf{Z}_t - \mathbf{Y}_t) \\
&-\frac{nT}{2} \log(\sigma_w^2) - \frac{T}{2} |S_w| - \frac{1}{2\sigma_w^2} \sum_{t=1}^{T} (\mathbf{Y}_t - \boldsymbol{\vartheta}_t)' S_w^{-1} (\mathbf{Y}_t - \boldsymbol{\vartheta}_t) \\
&+ \log\left(\pi(\rho, \boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_w^2, \phi_w)\right),
\end{aligned}$$

where $\pi(\rho, \boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_w^2, \phi_w)$ denotes the prior distribution, and $|S_w|$ denotes the determinant of S_w. We assume that *a priori* $\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\beta^2 I_p)$, and to have a flat prior distribution we take $\sigma_\beta^2 = 10^4$. The autoregressive coefficient ρ is also specified as the $N(0, 10^4)$ distribution, but restricted in the interval $I(0 < \rho < 1)$, so that this distribution is essentially flat. The inverse of the variance components, $1/\sigma_\epsilon^2$, $1/\sigma_w^2$, are assumed to follow $G(a, b)$ independently, where the distribution $G(a, b)$ has mean a/b. In our implementation, we take $a = 2$ and $b = 1$ to have a proper prior specification for each of these variance components, since improper prior distributions may lead to improper posterior distributions.

For the correlation decay parameter, ϕ_w we assume an independent uniform prior distribution in the interval $(0.001, 1)$. This corresponds to a value of spatial range between 3 and 3000 distance units (often taken as kilometers or miles). This prior distribution is appropriate for modeling air pollution data observed in a variety of study regions, e.g., a city where the maximum distance between any two locations is only a few kilometers or a substantial part of the eastern US where the maximum distance is approximately 3000 km. Clearly, the endpoints of the prior interval can be changed to accommodate the spatial range taking any meaningful value in a particular practical problem.

3. Prediction details

We first develop the methods for spatial interpolation of the air pollution levels at a new location $\mathbf{s}_0$ and any time t, $t = 1, \ldots, T$. Details for one-step ahead forecasting at time $t = T + 1$ are given at the forecasting subsection below. Spatial interpolation at location $\mathbf{s}_0$ and time t is based upon the predictive distribution of $Z(\mathbf{s}_0, t)$ given in the model Eqs (1) and (2). According to (1), $Z(\mathbf{s}_0, t)$, has the distribution:

$$Z(\mathbf{s}_0, t) \sim N\left(Y(\mathbf{s}_0, t),\ \sigma_\epsilon^2\right) \tag{6}$$

and

$$Y(\mathbf{s}_0, t) = \rho\, Y(\mathbf{s}_0, t-1) + x(\mathbf{s}_0, t)'\boldsymbol{\beta} + w(\mathbf{s}_0, t).$$

It is easy to see that $Y(\mathbf{s}_0, t)$ can only be sequentially determined using all the previous $Y(\mathbf{s}_0, t)$, including $Y(\mathbf{s}_0, 0)$, up to time t. Hence, we introduce the notation $\mathbf{Y}(\mathbf{s}, [t])$ to denote the vector $(Y(\mathbf{s}, 1), \ldots, Y(\mathbf{s}, t))'$ for $t \geq 1$. Note that a value of $Y(\mathbf{s}_0, 0)$ is required for this prediction problem. This value should be taken according to the prior distribution assumed for the initial condition on $\mathbf{Y}_0$. If, however, $\mathbf{Y}_0$ has been taken to be a fixed constant, then $Y(\mathbf{s}_0, 0)$ can also be taken as that same constant, as has been done in our illustration here.

The posterior predictive distribution of $Z(\mathbf{s}_0, t)$ is obtained by integrating over the unknown quantities in (6) with respect to the joint posterior distribution, i.e.,

$$\begin{aligned}\pi\left(Z(\mathbf{s}_0, t)|\mathbf{z}\right) = \int \pi\left(Z(\mathbf{s}_0, t)|Y(\mathbf{s}_0, [t]), \sigma_\epsilon^2\right)\, \pi\left(Y(\mathbf{s}_0, [t])|\boldsymbol{\theta}, \mathbf{v}\right) \\ \times\, \pi(\boldsymbol{\theta}, \mathbf{v}|\mathbf{z})\, dY(\mathbf{s}_0, [t])\, d\boldsymbol{\theta}\, d\mathbf{v}.\end{aligned} \tag{7}$$

When using MCMC methods to draw samples from the posterior, the predictive distribution (7) is sampled by composition. Draws from the posterior distribution $\pi(\boldsymbol{\theta}, \mathbf{v}|\mathbf{z})$ facilitates evaluation of the above integral, details are provided below.

We draw $Y(\mathbf{s}_0, t)$ from its conditional distribution given $\boldsymbol{\theta}, \mathbf{v}$ and $Y(\mathbf{s}_0, [t-1])$. Analogous to (5), we obtain for $t \geq 0$

$$\begin{pmatrix} Y(\mathbf{s}_0, t) \\ \mathbf{Y}_t \end{pmatrix} \sim N\left[\begin{pmatrix} \rho\, Y(\mathbf{s}_0, t-1) + x(\mathbf{s}_0, t)'\boldsymbol{\beta} \\ \rho\, \mathbf{Y}_{t-1} + X_t\boldsymbol{\beta} \end{pmatrix},\ \sigma_w^2 \begin{pmatrix} 1 & S_{w,12} \\ S_{w,21} & S_w \end{pmatrix}\right],$$

where $S_{w,12}$ is $1 \times n$ with the ith entry given by $\exp(-\phi_w d(\mathbf{s}_i, \mathbf{s}_0))$ and $S_{w,21} = S'_{w,12}$. Hence,

$$Y(\mathbf{s}_0, t)|\mathbf{Y}_t, \boldsymbol{\theta}, \mathbf{v} \sim N(\chi,\ \Lambda) \tag{8}$$

where $\Lambda = \sigma_w^2\left(1 - S_{w,12}S_w^{-1}S_{w,21}\right)$ and

$$\chi = \rho\, Y(\mathbf{s}_0, t-1) + x(\mathbf{s}_0, t)'\boldsymbol{\beta} + S_{w,12}S_w^{-1}\left(\mathbf{Y}_t - \rho\, \mathbf{Y}_{t-1} - X_t\boldsymbol{\beta}\right).$$

In summary, we implement the following algorithm to predict $Z(\mathbf{s}_0, t), t = 1, \ldots, T$.

1. Draw a sample $\boldsymbol{\theta}^{(j)}, \mathbf{v}^{(j)}, j \geq 1$ from the posterior distribution.
2. Draw $\mathbf{Y}^{(j)}(\mathbf{s}_0, [t])$ sequentially using (8). Note that the initial value $Y^{(j)}(\mathbf{s}_0, 0)$ is a constant for all $\mathbf{s}_0$ in our implementation.
3. Finally draw $Z^{(j)}(\mathbf{s}_0, t)$ from $N(Y^{(j)}(\mathbf{s}_0, t),\ \sigma_\epsilon^{2(j)})$.

The air pollution concentration on the original scale is the square of $Z^{(j)}(\mathbf{s}_0, t)$. If we want the predictions of the smooth pollution process without the nugget term, we simply omit the last step in the above algorithm and square the realizations $\mathbf{Y}^{(j)}(\mathbf{s}_0, t)$. We use the median of the MCMC samples and the lengths of the 95% intervals to summarize the predictions. The median as a summary measure preserves the one-to-one relationships between summaries for Y and Z, and for Y^2 and Z^2.

3.1. Calculating summaries

We now develop methodology for obtaining temporal summaries of air pollution. We illustrate by detailing methodologies for calculating the annual fourth highest ozone concentration at any site $\mathbf{s}_0$.

The true annual fourth highest daily maximum 8-hour average ozone concentration, denoted by $f(\mathbf{s}_0)$, is given by the fourth highest value of the series $Y^2(\mathbf{s}_0, 1), \ldots, Y^2(\mathbf{s}_0, T)$. (Note that we model ozone on the square-root scale.) At each MCMC iteration, j, we calculate $f^{(j)}(\mathbf{s}_0)$ and then the summaries of these posterior predictive realizations $f^{(j)}(\mathbf{s}_0)$ are used for predictions of the annual fourth highest daily maximum 8-hour average ozone concentration (and to obtain their uncertainties).

3.2. Forecasting

The one-step ahead Bayesian forecast at a location $\mathbf{s}_0$ is given by the posterior predictive distribution of $Z(\mathbf{s}_0, T+1)$, which is determined by $Y(\mathbf{s}_0, T+1)$. Note that using (8), we already have the conditional distribution of $Y(\mathbf{s}_0, T)$ given $\mathbf{Y}_t, \boldsymbol{\theta}$, and $\mathbf{v}$. We use model equation (2) to advance this conditional distribution one unit of time in future. The mean of the one-step ahead forecast distribution is given by $\rho\, Y(\mathbf{s}_0, T) + x(\mathbf{s}_0, T+1)'\boldsymbol{\beta}$, according to (2), and $Y(\mathbf{s}_0, T+1)$ should be equal to this if we are interested in forecasting the mean. If, however, we want to forecast an observation at location $\mathbf{s}_0$, we simulate $Y(\mathbf{s}_0, T+1)$ from the marginal distribution, which has the above mean and variance σ_w^2. We work with this marginal distribution rather than the conditional distribution like (8) above since conditioning with respect to the observed information (i.e., kriging) up to time T at the observation locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ has already been done to obtain $Y(\mathbf{s}_0, T)$, and at the future time $T+1$ there is no new available information to condition on except for the new values of the regressor, $x(\mathbf{s}_0, T+1)$. Then we follow the above algorithm to obtain the forecasts and their uncertainty estimates using ergodic averages of MCMC output.

4. An example

We illustrate with the daily maximum 8-hour average ozone concentration levels in 2006 observed in 117 monitoring sites in the three mid-western states namely, Illinois,

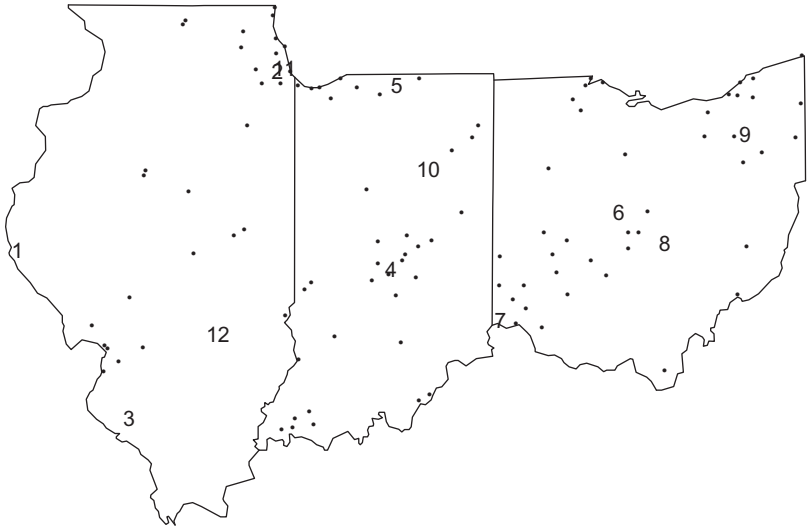

Fig. 1. A map of Illinois, Indiana, and Ohio with the 105 ozone monitoring sites plotted as points. The validation sites are labeled $1, \ldots, 12$.

Indiana, and Ohio. This study region (see Fig. 1) provides a good mix of developed industrial areas in Ohio and large cities like Chicago separated by vast rural areas. We use data from 12 randomly selected sites for validation purposes. The data from remaining 105 sites are used for model fitting.

Our analysis uses daily data for the $T = 153$ days in the high ozone season between May and September. This is a moderately large data set rich in both space and time with 16,065 observations ($= 105 \times 153$); 291 ($= 1.8\%$) of these are missing. The mean value is 47.62 parts per billion (ppb) and the range is 6.75–131.38 ppb. A sitewise boxplot (see Fig. 2) shows much spatial variation in the average levels between the sites. However, the variability within the sites is seen to be roughly constant, which can be explained by the fact that the daily observations are all based on eight hourly averages. A time series plot of the data for two randomly selected sites, provided in Fig. 3, shows high ozone values during the three hottest months of June, July, and August. The plot also shows the presence of moderate temporal dependence.

Following Sahu et al. (2009b), we include the output of a computer simulation model known as the CMAQ (Community Multiscale Air Quality), see *http://www.cmaq-model.org/*, as the single covariate in the model. The CMAQ model is based on emission inventories, meteorological information, and land use, and it produces average ozone concentration levels at each cell of a 12-km^2 grid covering the whole of the continental US, retrospectively, although there is a version of the model known as Eta-CMAQ that produces forecasts up to two days in advance. In this chapter, we use the retrospective daily maximum 8-hour average CMAQ ozone concentration for the grid cell covering the monitoring site. We provide a scatterplot of the ozone concentration values and the corresponding CMAQ values in Fig. 4. The plot shows a strong linear relationship between the two, but clearly CMAQ values are upwardly biased. This points to the need for a more accurate empirical model as is done here. Note that for this plot and also for modeling we have adopted the square-root scale. The spatial

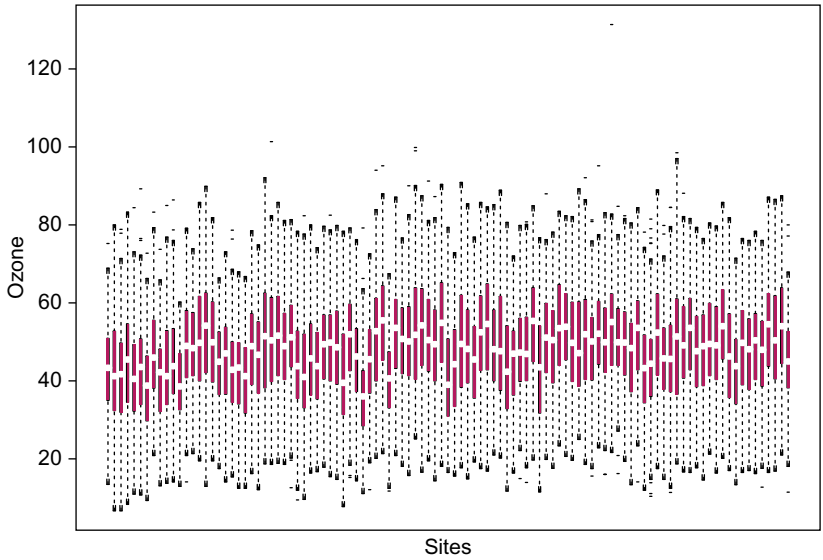

Fig. 2. Boxplots of the 153 daily maximum 8-hour ozone concentration levels in 2006 for each of the 105 sites located in Illinois and Indiana.

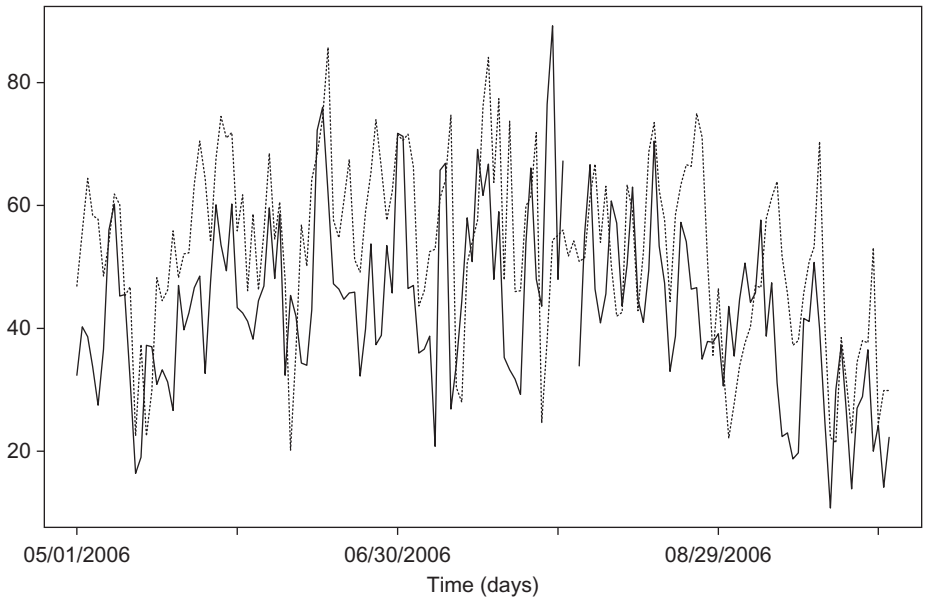

Fig. 3. Time series plot of daily ozone values from two randomly selected sites.

predictions at the unmonitored sites are performed using the CMAQ output at the corresponding grid cells. We have also attempted to include other meteorological covariates such as the daily maximum temperature, but none of those turned out to be significant in the presence of the CMAQ output.

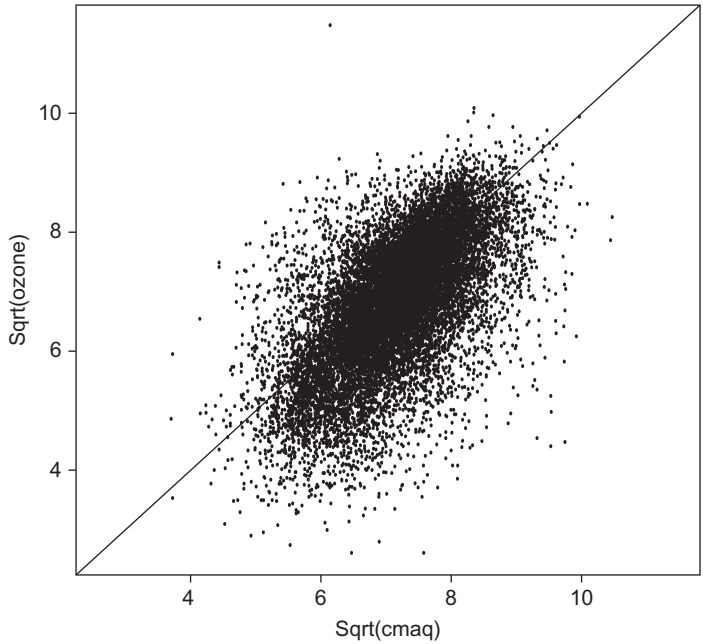

Fig. 4. A scatterplot of daily maximum 8-hour ozone concentration levels for the 105 sites in 2006 against the corresponding CMAQ values on the square-root scale.

In addition to the CMAQ output, we include an overall intercept β_0 in the model. Thus the mean of the true process $Y(\mathbf{s},t)$ is given by $\rho Y(\mathbf{s},t-1)+\beta_0+\beta x(\mathbf{s},t)$, where $x(\mathbf{s},t)$ denotes the CMAQ output at the grid cell that includes the location $\mathbf{s}$. The model also contains the two variance components σ_ϵ^2 and σ_w^2, and the spatial decay parameter, ϕ_w. Of course, all the $Y(\mathbf{s}_i,t)$ and the missing $Z(\mathbf{s}_i,t)$ are also need to be simultaneously estimated. We implement the Gibbs sampler with a Metropolis step for ϕ_w to simulate these parameters from their conditional distributions provided in the Appendix.

We tune the variance of the proposal distribution in the Metropolis step for the decay parameter ϕ_w to have a reasonable acceptance rate in the range $(0.15, 0.40)$. The tuning parameter finally adopted gave us an acceptance rate of 27.35% from 25,000 iterations. As is usual in MCMC computation, we have run the chains with many different starting values and monitored convergence by plotting traces of the parameters $\rho, \beta_0, \beta, \sigma_\epsilon^2, \sigma_w^2$, and ϕ_w. We have also examined the autocorrelation plots of these parameters and found those to be reasonable, i.e., the autocorrelations die down for moderate values of the lag parameter. There is, however, high crosscorrelation between the parameters σ_w^2 and ϕ_w and the MCMC chain mixes somewhat slowly because of this. This mixing problem occurs due to weak identifiability of the parameters and has been noted by many authors, see, e.g., Zhang (2004) and Stein (1999). The problem disappears if ϕ_w is not estimated and kept at a fixed value chosen by out of sample validations, see, e.g., Sahu et al. (2007). Here, we decide to sample ϕ_w for making inference using a large number of MCMC iterations, 20,000 after discarding the first 5000.

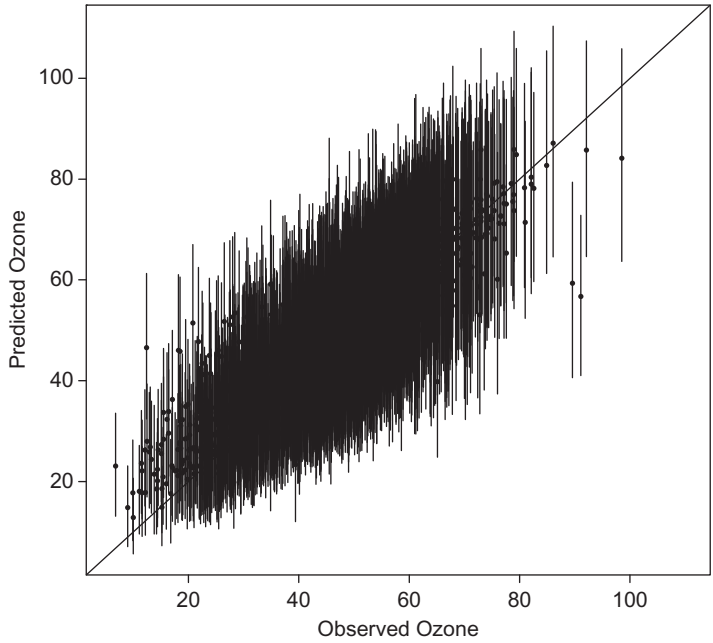

Fig. 5. A plot of the validation predictions against the observations along with 95% prediction intervals. The $y = x$ line is superimposed.

We use out of sample data from the 12 sites to validate the model. Out of the 1836 (12×153) validation ozone values, 32 are missing in our data. Figure 5 provides a plot of the 1804 out of sample predictions against the corresponding observations. The 95% prediction intervals are superimposed along with the $y = x$ line. The figure shows a very slight over and under prediction at the two ends of the ozone scale, otherwise, there is a good agreement between the observations and predictions. The nominal coverage of the 95% prediction intervals is 96.2%, which confirms the adequacy of the model. The validation mean square errors (VMSE) calculated for the 12 sites are between 8.91 to 114.68; the overall VMSE calculated using all the 1804 observations and their predictions is 38.02. These compare very favorably against CMAQ since the overall VMSE for CMAQ output is 144.18 and the sitewise CMAQ VMSEs range between 58.39 and 451.88. The overall VMSE value, 38.02, for the model-based predictions is also smaller than the same, 48.5, for a downscaler model recently proposed by Berrocal et al. (2011) for a similar data set.

The point and interval estimates of the model parameters are given in Table 1. We found moderate temporal dependence among successive day ozone concentrations (estimate of $\rho = 0.2687$). There is also strong spatial correlation, since the point estimate of ϕ_w is 0.0027 implying an approximate range of 1109 km. In addition to these strong spatial and temporal dependencies, the ozone concentrations also entertain the CMAQ output as a significant predictor, since the point estimate is 0.4976 and the 95% credible interval does not include zero. A direct interpretation of this estimate is difficult due to the square-root transformation used in modeling. Note that both autoregressive and the regression terms are significant predictors and hence their

Table 1
Estimation of the parameters. CI stands for equal-tailed credible intervals

	Mean	SD	95% CI
ρ	0.2687	0.0108	(0.2469, 0.2890)
β_0	1.4152	0.0667	(1.2885, 1.5485)
β	0.4976	0.0081	(0.4820, 0.5136)
σ_ϵ^2	0.2165	0.0042	(0.2085, 0.2248)
σ_w^2	0.4246	0.0229	(0.3848, 0.4738)
ϕ_w	0.0027	0.0002	(0.0024, 0.0031)

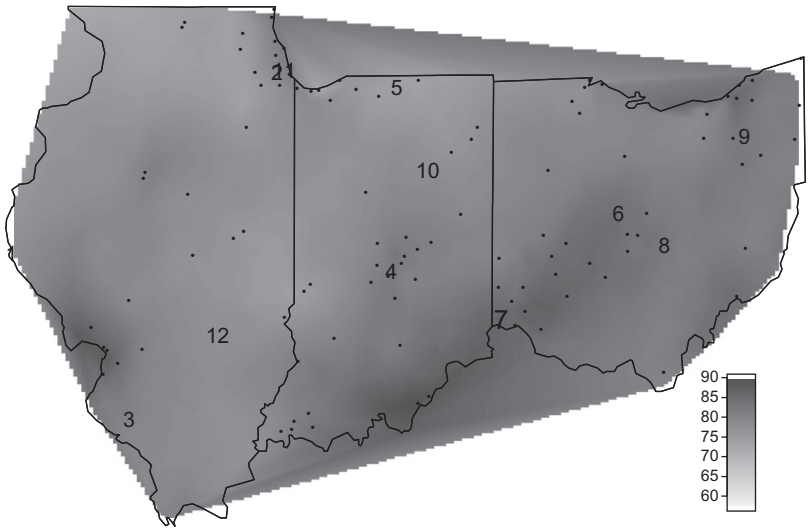

Fig. 6. Model-based predictions of the true annual fourth highest daily maximum 8-h average ozone levels in 2006. The 105 fitting sites are plotted as points and the validation sites are labeled $1, \ldots, 12$ as in Fig. 1.

inclusion will provide better model fitting and prediction and hence are retained in the model. Finally, the estimates of the variance components σ_ϵ^2 and σ_w^2 show that more variation is explained by the spatio-temporal effects than the pure error process $\epsilon(\mathbf{s}, t)$.

We now plot the annual fourth highest daily maximum 8-hour average true ozone values by linearly interpolating the predictions at the centers of the 900 randomly selected CMAQ grid cells in the study region (see Fig. 6). We find very good agreement among the predicted and observed maximum values. In fact, to quantify this with set aside data from the 12 validation sites, we provide the observed, the model predicted and the CMAQ output for the annual fourth highest daily maximum 8-hour average ozone concentration values in Table 2. The mean square error for the model-based

Table 2
Annual fourth highest daily maximum 8-hour average ozone concentrations in ppb units

Validation Site	Observed	Predicted	CMAQ
1	71.13	64.38	67.86
2	60.25	62.08	70.59
3	72.75	70.03	72.16
4	76.63	76.64	70.98
5	70.75	71.02	70.55
6	75.50	79.36	78.59
7	81.00	78.68	83.96
8	72.25	74.90	76.43
9	70.80	75.61	72.51
10	70.25	73.81	72.78
11	73.00	73.42	69.62
12	67.38	69.59	68.48

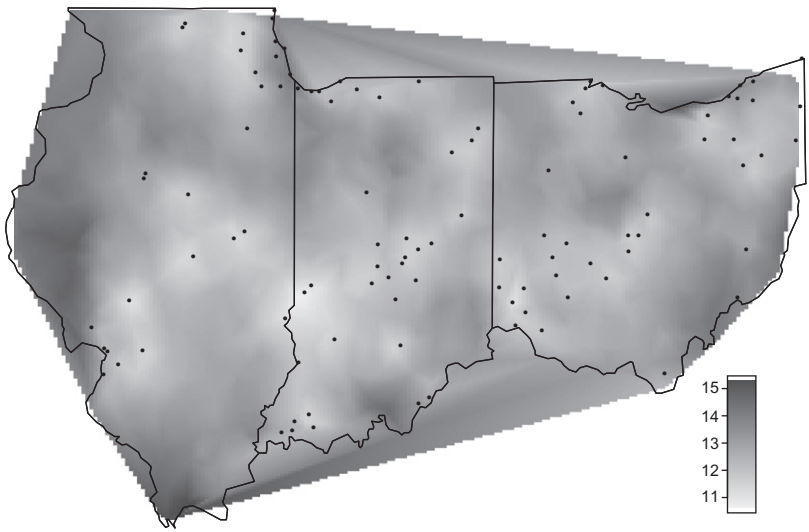

Fig. 7. Lengths of 95% prediction intervals for the true annual fourth highest daily maximum 8-hour ozone ozone levels in 2006.

predictions for these 12 validation sites is 10.4 while the same for the CMAQ output is 17.3. Thus, the model provides even more accurate predictions than the very accurate CMAQ output and the model is predicting the annual fourth highest value within a range of 3.2 ($= \sqrt{10.4}$) ppb on average. Figure 7 shows the uncertainties in the model predictions by providing a map of the lengths of the 95% prediction intervals. As expected, these intervals are larger in nonmonitored areas compared with monitored areas. No such uncertainty map is possible for the deterministic CMAQ output.

5. Further discussion

The modeling methods discussed in this chapter are suited for monitoring compliance with respect to air regulatory standards. High resolution spatial and fine scale temporal modeling allows inference on aggregated spatial (e.g. regional) and temporal summaries (e.g. annual). The Bayesian computation methods also enable accurate assessment of uncertainties in the aggregated summaries.

There are several other important areas of research in air pollution modeling. A number of papers are devoted to assessing exposure to air pollution and to fuse monitoring data with computer model output, see, e.g., Gelfand and Sahu (2010) for a recent review. Important modeling developments are also taking place in analyzing other pollutants such as sulfate and nitrate oxides. Deposition of these through precipitation is also of very much interest to researchers, see, e.g., Sahu et al. (2010) and the references therein. Forecasting of air pollution both for short-term and long-term periods also provide challenging statistical problems to the modeling community. Sahu et al. (2009a,b) develop methods for instantaneous forecasting of hourly and daily ozone levels.

Acknowledgment

The authors thank David Holland of the USEPA for many helpful comments, and also for providing the monitoring data and CMAQ model output used in this chapter.

Appendix: Conditional Distributions for Gibbs sampling

1. **Sampling Missing Data**. Any missing value, $Z(\mathbf{s},t)$ is to be sampled from $N(Y(\mathbf{s},t),\sigma_\epsilon^2)$, $t=1,\ldots,T$.
2. **Sampling σ_ϵ^2 and σ_w^2**. Straightforward calculation yields the following complete conditional distributions:

$$\frac{1}{\sigma_\epsilon^2} \sim G\left(\frac{nT}{2}+a, b+\tfrac{1}{2}\sum_{t=1}^{T}(\mathbf{Z}_t-\mathbf{Y}_t)'(\mathbf{Z}_t-\mathbf{Y}_t)\right),$$
$$\frac{1}{\sigma_w^2} \sim G\left(\frac{nT}{2}+a, b+\frac{1}{2}\sum_{t=1}^{T}(\mathbf{Y}_t-\boldsymbol{\vartheta}_t)'S_w^{-1}(\mathbf{Y}_t-\boldsymbol{\vartheta}_t)\right).$$

3. **Sampling $\mathbf{Y}_t$**. Let $Q_w=\Sigma_w^{-1}$. The full conditional distribution of $\mathbf{Y}_t$ is $N(\Lambda_t\boldsymbol{\chi}_t,\ \Lambda_t)$, where

Case 1: For $1 \le t < T-1$:

$$\Lambda_t^{-1} = \frac{I_n}{\sigma_\epsilon^2} + (1+\rho^2)Q_w,$$

$$\boldsymbol{\chi}_t = \frac{\mathbf{Z}_t}{\sigma_\epsilon^2} + Q_w\{\rho\mathbf{Y}_{t-1} + X_t\boldsymbol{\beta} + \rho\,(\mathbf{Y}_{t+1} - X_{t+1}\boldsymbol{\beta})\}.$$

Case 2: For $t = T$

$$\Lambda_t^{-1} = \frac{I_n}{\sigma_\epsilon^2} + Q_w,$$

$$\boldsymbol{\chi}_t = \frac{\mathbf{Z}_t}{\sigma_\epsilon^2} + Q_w\,(\rho \mathbf{Y}_{t-1} + X_t\boldsymbol{\beta}).$$

4. **Sampling** ρ. The full conditional distribution of ρ is $N(\Lambda\chi,\ \Lambda)$ where

$$\Lambda^{-1} = \sum_{t=1}^{T} \mathbf{Y}'_{t-1} Q_w \mathbf{Y}_{t-1} + 10^{-4},\ \chi = \sum_{t=1}^{T} \mathbf{Y}'_{t-1} Q_w (\mathbf{Y}_t - X_t\boldsymbol{\beta}),$$

restricted in the interval (0, 1).

5. **Sampling** $\boldsymbol{\beta}$. The full conditional distribution of $\boldsymbol{\beta}$ is $N(\Lambda\boldsymbol{\chi},\ \Lambda)$, where

$$\Lambda^{-1} = \sum_{t=1}^{T} X'_t Q_w X_t + \Sigma_\beta^{-1}, \text{ and}$$

$$\boldsymbol{\chi} = \sum_{t=1}^{T} X'_t Q_w (\mathbf{Y}_t - \rho \mathbf{Y}_{t-1}).$$

6. **Sampling** ϕ_w. The full conditional distribution of ϕ_w is nonstandard and must be calculated from the prior and likelihood terms involving ϕ_w and is given by:

$$\log(\pi(\phi_w|\cdots) = -\frac{1}{2}|S_w| - \frac{1}{2\sigma_w^2}\sum_{t=1}^{T}(\mathbf{Y}_t - \boldsymbol{\vartheta}_t)' S_w^{-1}(\mathbf{Y}_t - \boldsymbol{\vartheta}_t) + \log(\pi(\phi_w))$$

up to a normalizing constant, where $\cdots$ denotes all the data and parameters except for ϕ_w. We adopt a Metropolis–Hastings step to obtain sample from this full conditional distribution as follows. Let $\phi_w^{(p)}$ denote a sample from a proposal distribution $q\left(\phi_w^{(p)}|\phi_w^{(c)}\right)$ where the current value is $\phi_w^{(c)}$. The sampled value, $\phi_w^{(p)}$ is accepted with probability

$$\alpha\left(\phi_w^{(p)},\ \phi_w^{(c)}\right) = \min\left\{1,\ \frac{\pi(\phi_w^{(p)}|\cdots)\, q\left(\phi_w^{(c)}|\phi_w^{(p)}\right)}{\pi(\phi_w^{(c)}|\cdots)\, q\left(\phi_w^{(p)}|\phi_w^{(c)}\right)}\right\}.$$

This acceptance probability simplifies considerably when $q\left(\phi_w^{(p)}|\phi_w^{(c)}\right)$ is taken to be symmetric in its arguments $\phi_w^{(p)}$ and $\phi_w^{(c)}$, i.e., when $q\left(\phi_w^{(p)}|\phi_w^{(c)}\right) = q\left(\phi_w^{(c)}|\phi_w^{(p)}\right)$. In this case, the ratio of densities in the acceptance probability is simply calculated by the ratio $\pi\left(\phi_w^{(p)}|\cdots\right)/\pi(\phi_w^{(c)}|\cdots\Big)$. The resulting algorithm is known as the Metropolis algorithm.

We implement the Metropolis algorithm by taking the proposal distribution as the normal distribution with the mean at the current value and the variance σ_p^2, which we tune to have an acceptance rate between 15% and 40%, see Gelman et al. (1996) for theoretical justifications. Moreover, we implement the Metropolis algorithm on the log-scale for ϕ_w, i.e., we work with the density of $\log(\phi_w)$ instead of ϕ_w since the support of the normal proposal distribution is the real line. Keeping ϕ_w within a range is trivial since any proposal value outside the range is rejected forthwith.

References

Abramowitz, M., Stegun, I.A., 1965. Handbook of Mathematical Functions. Dover, New York.

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC, Boca Raton.

Berrocal, V.J., Gelfand, A.E., Holland, D.M., 2010. A spatio-temporal downscaler for output from numerical models. J. Agric. Biol. Environ. Stat. 15, 176–197.

Berrocal, V.J., Gelfand, A.E., Holland, D.M., 2011. Space-time data fusion under error in computer model output: an application to modeling air quality. Technical Report. Duke University.

Brown, P.J., Le, N.D., Zidek, J.V., 1994. Multivariate spatial interpolation and exposure to air pollutants. Can. J. Stat. 22, 489–510.

Carroll, R.J., Chen, R., George, E.I., Li, T.H., Newton, H.J., Schmiediche, H., et al., 1997. Ozone exposure and population density in Harris County, Texas. J. Am. Stat. Assoc. 92, 392–404.

Cocchi, D., Fabrizi, E., Trivisano, C., 2005. A stratified model for the assessment of meteorologically adjusted trends of surface ozone. Environ.dtins, Ecol. Stat. 12, 1195–1208.

Cocchi, D., Greco, F., Trivisano, C., 2007. Hierarchical space-time modelling of PM10 pollution. Atmos. Environ. 41, 532–542.

Cox, W.M., Chu, S.H., 1992. Meteorologically adjusted trends in urban areas, a probabilistic approach. Atmos. Environ. 27, 425–434.

Cressie, N., 1994. Comment on "An approach to statistical spatial-temporal modeling of meteorological fields" by M.S. Handcock and J.R. Wallis. J. Am. Stat. Assoc. 89, 379–382.

Cressie, N., Shi, T., Kang, E.L., 2010. Fixed rank filtering for spatio-temporal data. J. Comput. Graph. Stat. 19, 724–745.

Cressie, N., Kaiser, M.S., Daniels, M.J., Aldworth, J., Lee, J., Lahiri, S.N., et al., 1999. Spatial analysis of particulate matter in an urban environment. In: Gmez-Hernndez, J., Soares, A., Froidevaux, R. (Eds.), GeoEnv II: Geostatistics for Environmental Applications. Dordrecht, Kluwer, pp. 41–52.

Cressie, N., Wikle, C.K., 2011. Statistics for Spatio-Temporal Data. John Wiley & Sons, Hoboken.

Dou, Y., Le, N.D., Zidek, J.V., 2010. Modeling hourly ozone concentration fields. Ann. Appl. Stat. 4, 1183–1213.

Gelman, A., Roberts, G.O., Gilks, W.R., 1996. Efficient Metropolis jumping rules. In: Bernardo, J.O., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 5. Oxford, Oxford University Press, pp. 599–607.

Gelfand, A.E., Banerjee, S., Gamerman, D., 2005. Spatial process modeling for univariate and multivariate dynamic spatial data. Environmetrics 16, 465–479.

Gelfand, A.E., Sahu, S.K., 2010. Combining monitoring data and computer model output in assessing environmental exposure. In: OHagan, A., West, M. (Eds.), Handbook of Applied Bayesian Analysis. Oxford University Press, Oxford, pp. 482–510.

Goodall, C., Mardia, K.V., 1994. Challenges in multivariate spatio-temporal modeling. In: Proceedings of the XVIIth International Biometric Conference. Hamilton, Ontario, Canada, 8–12 August 1994, pp. 1–17.

Guttorp, P., Meiring, W., Sampson, P.D., 1994. A space-time analysis of ground-level ozone data. Environmetrics, 5, 241–254.

Huang, L.S., Smith, R.L., 1999. Meteorologically-dependent trends in urban ozone. Environmetrics 10, 103–118.

Huerta, G., Sanso, B., Stroud, J.R., 2004. A spatiotemporal model for Mexico City ozone levels. J. Roy. Stat. Soc. C, 53, 231–248.

Kibria, B.M.G., Sun, L., Zidek, J.V., Le, N.D., 2002. Bayesian spatial prediction of random space-time fields with application to mapping PM2.5 exposure. J. Am. Stat. Assoc. 97, 112–124.

Kyriakidis, P.C., Journal, A.G., 1999. Geostatistical space-time models: A review. Math. Geol. 31, 651–684.

Mardia K.V., Goodall C., Redfern E.J., Alonso F.J., 1998. The Kriged Kalman filter (with discussion). Test 7, 217–252.

McMillan, N., Bortnick, S.M., Irwin, M.E. Berliner, M., 2005. A hierarchical Bayesian model to estimate and forecast ozone through space and time. Atmos. Environ. 39, 1373–1382.

Pollice, A., Lasinio, G.J., 2010. Spatiotemporal analysis of PM10 concentration over the Taranto area. Environ. Monit. Assess. 162, 177–190.

Porter, P.S., Rao, S.T., Zurbenko, I.G., Dunker, A.M., Wolff, G.T., 2001. Ozone air quality over North America: part II – an analysis of trend detection and attribution techniques. J. Air Waste Manag. Assoc. 51, 283–306.

Sahu, S.K., Mardia, K.V., 2005. A Bayesian Kriged-Kalman model for short-term forecasting of air pollution levels. J. R. Stat. Soc. Ser. C 54, 223–244.

Sahu, S.K., Gelfand, A.E., Holland, D.M., 2006. Spatio-temporal modeling of fine particulate matter. J. Agric. Biol. Environ. Stat. 11, 61–86.

Sahu, S.K., Gelfand, A.E. Holland, D.M., 2007. High resolution space-time ozone modeling for assessing trends. J. Am. Stat. Assoc. 102, 1221–1234.

Sahu, S.K., Gelfand, A.E. Holland, D.M., 2010. Fusing point and areal space-time data with application to wet deposition. J. R. Stat. Soc. Ser. C 59, 77–103.

Sahu, S.K., Yip, S., Holland, D.M., 2009a. A fast Bayesian method for updating and forecasting hourly ozone levels. Environ. Ecol. Stat. 18, 185–207. doi:10.1007/s10651-009-0127-y.

Sahu, S.K., Yip, S., Holland, D.M., 2009b. Improved space-time forecasting of next day ozone concentrations in the eastern US. Atmos. Environ. 43, 494–501. doi:10.1016/j.atmosenv.2008.10.028.

Stein, M., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer-Verlag, New York.

Stroud, J.R., Müller, P. Sansó, B., 2001. Dynamic models for Spatio-temporal data. J. R. Stat. Soc. B 63, 673–689.

Shaddick, G., Wakefield, J., 2002. Modelling daily multivariate pollutant data at multiple sites. J. R. Stat. Soc. Ser. C 51, 351–372.

Smith, R.L., Kolenikov, S., Cox, L.H., 2003. Spatio-Temporal modelling of $PM_{2.5}$ data with missing values. J. Geophys. Res. Atmos. 108, NO. D24, 9004, doi:10.1029/2002JD002914.

Sun L., Zidek, J.V., Le, N.D., Ozkaynak, H., 2000. Interpolating Vancouver's daily ambient PM10 field. Environmetrics 11, 651–663.

Thompson, M.L., Reynolds, J., Cox, L.H., Guttorp, P., Sampson, P.D., 2001. A review of statistical methods for the meteorological adjustment of tropospheric ozone. Atmos. Environ. 35, 617–630.

Wikle, C.K., 2003. Hierarchical models in environmental science. Int. Stat. Rev. 71, 181–199.

Wikle, C.K., Cressie, N., 1999. A dimension-reduced approach to space-time Kalman filtering. Biometrika 86, 815–829.

Zhang, H., 2004. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. J. Am. Stat. Assoc. 99, 250–261.

Zhu, L., Carlin, B.P., Gelfand, A.E., 2003. Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in Atlanta. Environmetrics 14, 537–557.

Zidek, J.V., Sun, L., Le, N., Ozkaynak, H., 2002. Contending with space-time interaction in the spatial prediction of pollution: Vancouver's hourly ambient PM_{10} field. Environmetrics 13, 595–613.

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00017-X

Karhunen–Loéve Expansion of Temporal and Spatio-Temporal Processes

Lara Fontanella and Luigi Ippoliti

Department of Economics, University G. d'Annunzio, Viale Pindaro 42, 65127 Pescara, Italy

Abstract

In this chapter, we describe the Karhunen–Loéve expansion (KLE) of temporal and spatio-temporal processes. KLE is one of the most frequently used statistical techniques for data mining of continuous stochastic processes. In its discrete formulation, it is simply empirical orthogonal function (EOF) analysis or principal component analysis (PCA). Hence, both KLE and EOF are useful to achieve efficient dimension reduction of huge data sets.

There are several references that discuss the use of EOF, or KLE, in the climatological sciences. Because their applications are not only restricted to atmospheric sciences, the aim of this chapter is to frame KLE in a more general setting by addressing various recent developments.

Keywords: temporal processes, spatial processes, spatio-temporal processes, biomedical time series, Karhunen–Loéve expansion, empirical orthogonal functions, principal component analysis.

1. Introduction

Temporal and spatio-temporal models have received widespread popularity and have been largely developed through applications in many scientific fields such as engineering, economics, environmental sciences, climate prediction, and meteorology. More recent activities in the area also include tracking, functional MRI, and health data analysis.

The theory and practice of time series analysis have developed rapidly since the publication of the book by Box and Jenkins (1970). Development and research in the

spatio-temporal area started only in the last 20 years, when management and manipulation of data, relating to both spatial and temporal changes, were recognized as an indispensable assignment. In the nineties, several researchers independently began looking at the potential of having a dynamic temporal aspect in space–time statistical modeling. However, until recently, there has not been a theory of spatio-temporal processes separate from the already well-established theories of spatial statistics and time series analysis. The books by Banerjee et al. (2004) and Sherman (2011) provide a good starting point for researchers in this area.

Motivated by different applications, various modeling strategies have been adopted; the choice of the approach is generally dictated by the objective of the study, whether it be obtaining forecasts, estimating trends, or increasing the scientific understanding of the underlying mechanisms.

A key challenge of many research studies is the extraction of information from large temporal or spatio-temporal data sets now available. These data sets often comprise observations of extremely complicated underlying processes. Hence, methods of analysis must be able to account for multiscale dynamical variability across different dynamical variables in space and time, account for various sources of error, and provide efficient dimension reduction. Scientists have developed or borrowed and refined many descriptive statistical techniques that aid in the summary and interpretation of these data. The focus here is on the Karhunen–Loéve (KL) expansion, which is one of the most frequently used statistical techniques for data mining of continuous stochastic processes. Note that, in its discrete formulation, KL analysis is simply empirical orthogonal function (EOF) analysis or principal component analysis (PCA). Since the introduction by Lorenz (1956), we can find an extensive use of EOFs in the atmospheric sciences. For example, they have been used for describing climate, comparing simulations of general circulation models, developing regression forecast techniques, weather classification, map typing, the interpretation of geophysical fields, and the simulation of random fields, particularly nonhomogeneous processes. For a general discussion of EOFs in meteorology, see, for example, Craddock (1973).

The interested reader should note that there are several excellent reference books (see, for example, Jolliffe (2002) and von Storch and Zwiers (1999)) and review papers (Hannachi et al., 2007; Monahan et al., 2009) that discuss the use of EOF or KL in the climatological sciences. However, the applications of KL and EOF are restricted to atmospheric sciences, and thus, the subject has not been systematically reviewed in a more general literature to address the various recent developments. This chapter is a contribution to fill in this gap, but is by no means exhaustive.

The plan of this chapter is as follows. In Section 2, we introduce the theory of Karhunen–Loéve analysis for one-dimensional continuous stochastic process; we show that the analysis consists of two complementary stages, and we describe them through EOF analysis. In Section 3, we describe a multiresolution version of KL and show its usefulness for the analysis of biomedical signals. Sections 4 and 5, instead, extend the KL theory to coupled one-dimensional processes and spatio-temporal processes, respectively. Section 6 concludes the chapter with a discussion.

2. Karhunen–Loéve expansion of one-dimensional processes

A random process can be represented as a series expansion involving a complete set of deterministic functions with corresponding random coefficients. There are several such

series expansions that are widely in use. A commonly used one is the Fourier series in which the coefficients are real numbers and the expansion basis consists of sinusoidal functions. Zhang and Ellingwood (1994) proposed another orthogonal series expansion using Legendre polynomials as the deterministic basis function. However, the random coefficients in the expansion are correlated random variables. Other polynomials have also been used (Li and Der-Kiureghian, 1993). The use of Karhunen–Loéve expansion with orthogonal deterministic basis functions and uncorrelated random coefficients has generated interest because of its biorthogonal property, that is, both the deterministic basis functions and the corresponding random coefficients are orthogonal. This allows for the optimal encapsulation of the information contained in the random process into a set of discrete uncorrelated random variables (Ghanem and Spanos, 1991).

Let $Y(t)$ be a random process defined on a probability space $(\Omega; A; P)$ and indexed on a bounded domain T. Assume that $Y(t)$ has a mean $\mu(t)$, such that $X(t) = Y(t) - \mu(t)$ is a zero-mean process with finite variance, $E\left[X(t)^2\right]$, that is bounded for all $t \in T$, and continuous covariance function, $R(t,t') = E\left[X(t)X(t')\right]$. The mean-value function $\mu(t)$ is usually unknown but can be estimated from the data when some priori knowledge is available concerning the functional form of $\mu(t)$. For example, when $Y(t)$ is nonstationary, $\mu(t)$ may be well approximated by a polynomial of degree $k > 0$.

It is well known (see, for example, Loève (1978), Shorack and Wellner (1986) and Karhunen (1947)) that there exist constants, $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$, together with continuous functions $\phi_1(t), \phi_2(t), \ldots$, such that the following properties are fulfilled:

P.1 The set $\{\phi_i,\ i \geq 1\}$ forms a complete orthogonal system in the space of a square integrable function $L_2(T)$, that is,

$$\int_T \phi_i(t)\phi_j(t)dt = \delta_{ij},$$

where δ_{ij} is the Kronecker-delta function.

P.2 The set $\{(\lambda_i, \phi_i),\ i \geq 1\}$ forms a complete set of solutions of the Fredholm-type equation in (λ, ϕ)

$$\int_T R(t,t')\phi_i(t)dt = \lambda_i\phi_i(t') \quad \text{and} \quad \int_T \phi_i^2(t)dt = 1. \tag{1}$$

P.3 From Mercer's theorem (Riesz and Sz-Nagy, 1955), we have the following spectral or eigen-decomposition

$$R(t,t') = \sum_{i=1}^{\infty} \lambda_i\phi_i(t)\phi_i(t'), \tag{2}$$

which is a series absolutely and uniformly convergent in (t,t').

P.4 There exists a sequence, $\{z_i,\ i \geq 1\}$, of zero-mean uncorrelated random variables with variance $E\left[z_i^2\right] = \sigma_{z_i}^2 = \lambda_i$, given by the inner product

$$z_i = \int_T X(t)\phi_i(t)dt, \tag{3}$$

such that the following KL expansion holds

$$Y(t) = \mu(t) + \sum_{i=1}^{\infty} z_i \, \phi_i(t). \tag{4}$$

Note that if $Y(t)$ is restricted to a Gaussian process, then the z_i are also uncorrelated Gaussian random variables. It is also customary to standardize the z_i to unit variance, but in what follows we assume that they are unstandardized unless specified otherwise.

The series expansion in (4), which is known to converge in the mean square sense for any distribution of $Y(t)$, is referred to as the KL expansion (KLE) and provides a second moment characterization in terms of uncorrelated random variables and deterministic orthogonal functions.

For practical implementation, the series (2) and (4) are approximated by a finite number of terms, say M, giving

$$R(t, t') = \sum_{i=1}^{M} \lambda_i \phi_i(t) \phi_i(t')$$

and

$$Y(t) = \mu(t) + \sum_{i=1}^{M} z_i \, \phi_i(t). \tag{5}$$

In the study by Grenander (1976) and Ghanem and Spanos (1991), it is shown that this truncated series is optimal. That is, ordering the terms of the expansion in decreasing order of the variances, λ_i, of the coefficients, z_i, the KLE gives an optimal expansion in the sense that the series truncated at any point minimizes the integrated mean square error between the actual and approximated random function. In other words, if we approximate the random process in terms of M basis functions, the optimal basis functions for the truncated expansion correspond to the eigenvectors of the covariance matrix, $\mathbf{R}$, with the M largest eigenvalues. Hence, for any other set of coefficients, $d_i \neq z_i$, we have the inequality

$$\int_T \left[X(t) - \sum_{i=1}^{M} z_i \, \phi_i(t) \right]^2 dt \leq \int_T \left[X(t) - \sum_{i=1}^{M} d_i \, \phi_i(t) \right]^2 dt.$$

Also, the expansion (2) minimizes the entropy measure

$$I_\lambda = \sum_{i=1}^{M} \lambda_i \ln(\lambda_i).$$

Thus, the KLE is optimal in the sense of simultaneously minimizing a mean-squared criterion and maximizing information content.

Depending on the application, the Karhunen–Loéve analysis is usually used in a diagnostic mode to find principal (in terms of explanation of variance) temporal structures and to reduce the dimension in large data sets while simultaneously reducing noise. Hence, an objective of KL analysis is to make a decomposition of the original series into the sum of a small number of independent and interpretable components

$$Y(t) = T(t) + C(t) + S(t) + E(t),$$

where $T(t)$ is a polynomial trend, $C(t)$ is a cycle, generally possessing variable amplitude, and $S(t)$ denotes seasonal or periodic within-year movements. For economic time series, $T(t)$, $C(t)$, and $S(t)$ are typically viewed as long-, medium-, and short-term movements, respectively. The $E(t)$ is a structureless noise assumed to be a zero mean stationary random process with variance σ_e^2, uncorrelated with the signal component,

$$V(t) = T(t) + C(t) + S(t), \tag{6}$$

which is an unobservable process "smoother" than $Y(t)$.

Without specifying any parametric model for the observed time series, the Karhunen–Loéve analysis is thus useful for solving the following problems: (1) finding trends of different resolution, (2) smoothing, (3) extraction of seasonality components, (4) simultaneous extraction of cycles with small and large periods, (5) extraction of periodicities with varying amplitudes, (6) simultaneous extraction of complex trends and periodicities, (7) finding structures in short time series, and (8) change-point detection.

The Karhunen–Loéve technique consists of two complementary stages: *decomposition* and *reconstruction*, both of which include two separate steps. These will be discussed in the next sections.

2.1. Decomposing discrete time series

2.1.1. Embedding

Consider the zero-mean time series $X(t)$. The embedding step is associated with the construction of the so-called *trajectory* matrix. Several approaches can be followed to define this matrix, and some examples can be found in the study by Basilevsky and Hum (1979), Golyandina et al. (2001), and Hannachi et al. (2007). Here, we follow Fontanella et al. (2010) and define the trajectory matrix as N time delayed and decimated copies of the observed time series, $\mathbf{x} = \big(x(1), \ldots, x(n)\big)'$. Thus, embedding can be regarded as a projection of the one-dimensional series onto an N-dimensional hyperspace. Without loss of generality, it is assumed that both the length of the signal, n, and the number of its time delayed copies, N, are power of two; that is, $n = 2^J$ and $N = 2^K$. Hence, the $(2^{J-K+1} \times N)$ trajectory matrix $\mathbf{X}$ has generic $X(i, j)$ element given by

$$X(i, j) = 2^{-1/2} x\left(j + 2^{K-1}(i-1)\right); \quad i = 1, \ldots, 2^{J-K+1} \quad j = 1, \ldots, N. \tag{7}$$

For example, assuming $J=4$ and $K=2$, the $\mathbf{X}$ matrix shows the following structure

$$\mathbf{X}=\frac{1}{\sqrt{2}}\begin{bmatrix} x(1) & x(2) & x(3) & x(4) \\ x(3) & x(4) & x(5) & x(6) \\ x(5) & x(6) & x(7) & x(8) \\ x(7) & x(8) & x(9) & x(10) \\ x(9) & x(10) & x(11) & x(12) \\ x(11) & x(12) & x(13) & x(14) \\ x(13) & x(14) & x(15) & x(16) \\ x(15) & x(16) & x(1) & x(2) \end{bmatrix},$$

where, for simplicity, "circular boundary conditions" are imposed on the signal but other solutions can be considered. Circular boundary conditions are easily dealt with and have been found useful in biomedical applications (see, for example, Fontanella et al. (2010)). We also note that the N delayed copies of the original signal are decimated with a sampling rate of 2^{K-1} steps and that the last 2^{K-1} elements of each row vector equals the first 2^{K-1} entries of its immediate neighbor, with each observation thus repeated twice. Then, to ensure that data matrix preserves the energy of the original signal, Eq. (7) multiplies each elements in $\mathbf{X}$ by $1/\sqrt{2}$. This, in fact, ensures that $\|\mathbf{X}\|_F=(\mathbf{x}'\mathbf{x})^{1/2}$, where $\|\cdot\|_F$ is the Frobenius norm.

2.1.2. Eigenvalue decomposition

The second step deals with the eigenvalue decomposition of the estimated covariance matrix

$$\widehat{\mathbf{R}}=2^{-(J-K+1)}\mathbf{X}'\mathbf{X}.$$

Note that the estimator above is largely motivated by the requirements of the principal components model and intuitive appeal, rather than because they are best in any known sense. For a different approach, see, for example, Section 5 and the discussion in the study by Basilevsky and Hum (1979). Denote with $\hat{\lambda}_1,\ldots,\hat{\lambda}_N$ the eigenvalues of $\widehat{\mathbf{R}}$ sorted in decreasing order (i.e., $\hat{\lambda}_1\geq\hat{\lambda}_2,\ldots,\hat{\lambda}_N\geq 0$) and with $\hat{\boldsymbol{\phi}}_1,\ldots,\hat{\boldsymbol{\phi}}_N$ the associated eigenvectors. Then it follows that, $\sum_{i=1}^{N}\hat{\lambda}_i=tr\left(\widehat{\mathbf{R}}\right)=2^{K-1}\hat{\sigma}_x^2$, where $\hat{\sigma}_x^2=2^{-J}(\mathbf{x}'\mathbf{x})$.

Let $\hat{\mathbf{z}}_i=\mathbf{X}\hat{\boldsymbol{\phi}}_i$, then the trajectory matrix can be written as

$$\mathbf{X}=\mathbf{X}_1+\mathbf{X}_2+\cdots+\mathbf{X}_N, \tag{8}$$

where $\mathbf{X}_i=\hat{\mathbf{z}}_i\hat{\boldsymbol{\phi}}_i^{'}, i=1,\ldots,N$. The matrices $\mathbf{X}_i$ have rank 1 and are therefore elementary matrices; the $\hat{\boldsymbol{\phi}}_i$ are also known as "empirical orthogonal functions" or simply EOFs, whereas the $\hat{\mathbf{z}}_i$ are the principal components or the right eigenvectors of the trajectory matrix. The collection $\left(\sqrt{\hat{\lambda}_i},\hat{\boldsymbol{\phi}}_i,\hat{\mathbf{z}}_i\right)$ is called the ith *eigentriple* of the matrix $\mathbf{X}$, $\sqrt{\hat{\lambda}_i}$ are the singular values of the matrix $\mathbf{X}$, and the set $\left\{\sqrt{\hat{\lambda}_i}\right\}$ is called the spectrum of $\mathbf{X}$.

Note that the rows and columns of the trajectory matrix $\mathbf{X}$ are subseries of the original time series. Therefore, the left eigenvectors, $\hat{\boldsymbol{\phi}}_i$, and principal components, $\hat{\mathbf{z}}_i$, also have a temporal structure and hence can also be regarded as time series.

2.1.3. The window length

The window length, $N = 2^K$, is the only parameter one needs to choose at this stage. Selection of the proper window length depends on the problem in hand and on preliminarily information about the time series. For a deeper discussion on this point, the reader is referred to Golyandina et al. (2001). In general, if we know that the time series may have a periodic component with an integer period (e.g., if this component is a seasonal component), then to get better separability of this periodic component, it is advisable to take the window length proportional to that period. It is also worth noting that the choice of K determines the maximum temporal lag within the autocovariance matrix $\widehat{\mathbf{R}}$, and the support length of the estimated basis functions $\phi_i(t), i = 1, \ldots, N$. In general, the larger is K, the smaller are the size fluctuation values produced by the KLE trasform. For example, for two values of K, say K_1 and K_2, with $K_1 > K_2$, if the goal is compression of a signal, then a KL based on K_1 will generally perform better than KL based on K_2; on the other hand, if the goal is identifying features of the signal that are related to turning points in its graph, then the KL based on K_2 can identify the locations of these turning points more clearly.

2.2. Reconstruction

2.2.1. Aggregation

In practice, one cannot always identify the time series components with corresponding unique orthogonal variables (time functions) $\mathbf{z}_i$. For example, the cycle effect $C(t)$ may be represented by several independent time functions. The grouping step corresponds to aggregating the elementary matrices $\mathbf{X}_i$ into m groups and summing the matrices within each group. Hence, for a set of p indices, $\{j_1, \ldots, j_p\}$, we define the aggregated matrix, $\tilde{\mathbf{X}}_g$, as

$$\tilde{\mathbf{X}}_{g_k} = \mathbf{X}_{j_1} + \mathbf{X}_{j_2} + \cdots + \mathbf{X}_{j_p}, \quad k = 1, \ldots, m,$$

so that the clustering of the set of indices, $i = \{1, \ldots, N\}$, into m disjoint subsets, $G = \{g_1, \ldots, g_m\}$, corresponds to the representation

$$\mathbf{X} = \tilde{\mathbf{X}}_{g_1} + \tilde{\mathbf{X}}_{g_2} + \cdots + \tilde{\mathbf{X}}_{g_m}.$$

The procedure of choosing the sets $g_1, \ldots, g_m$, is called the eigentriple clustering. For a given group, g_k, the contribution of the component $\tilde{\mathbf{X}}_{g_k}$ into the expansion (8) is measured by the share of the corresponding eigenvalues, $\sum_{i \in g_k} \hat{\lambda}_i / \sum_{j=1}^{N} \hat{\lambda}_j$.

For purposes of interpretation, however, one may prefer to represent similar time behavior by a single component. Therefore, our application of KL analysis to time series data does not escape the identification problem, which is well known in the factor analysis literature. Basilevsky and Hum (1979) suggest the following two-stage procedure by which the $\mathbf{z}_i$ can be aggregated into $T(t)$, $C(t)$, and $S(t)$:

1. Plot the random variables $\mathbf{z}_i$ against time to allow a visual inspection of their temporal pattern. Usually the first few $\mathbf{z}_i$, which correspond to large eigenvalues, will reveal the presence of trend, cycle, and seasonality, if these terms exist at all in $X(t)$. Pairwise scatterplots of the time series $\mathbf{z}_i$ and $\mathbf{z}_j$, for $i \neq j$, may also

help one to visually identify those eigentriples that corresponds to the harmonic components of the series;

2. When more than one $\mathbf{z}_i$ exhibits similar time behavior, for example, a cyclic with a given period, a χ^2 criterion can be used to test which of the $\mathbf{z}_i$ are to be clustered into a common cyclic term $C(t)$ and which are to be retained as distinct cycles. For practical purposes, it is sufficient to test for equality of the eigenvalues alone, and this may be done conveniently by Anderson's (1963) large sample statistic

$$\chi^2 = -c\sum_{i=1}^{r}\ln(\lambda_i) + c\,r\ln\left(\sum_{i=1}^{r}\frac{\lambda_i}{r}\right)$$

with $\frac{1}{2}r(r+1)-1$ degrees of freedom, where r is the number of roots to be tested.

In addition, further strategies may be considered for the identification of the eigentriples of the matrix $\mathbf{X}$. For example, it happens in practice that the singular values of the two eigentriples of a harmonic series are often very close to each other, thus, simplifying their clustering. The periodogram analysis of the series $\mathbf{z}_i$ may also reveal important features and can help in making the clustering. In fact, as shown in the study by Fontanella et al. (2010), this approach has been found helpful in identifying specific latent components of biomedical time series. Coli et al. (2005) also show that investigation of the spectral features of the basis functions is helpful for the estimation of short- and long-memory parameters of $ARFIMA(p,d,q)$ models.

2.2.2. Averaging

The signal, $X(t)$, is reconstructed exactly by averaging the elements that are repeated twice:

$$x(t) = \sqrt{2}\,mean\big\{X(i,j) : 1 \le i \le 2^{J-K+1}, \quad 1 \le j \le N,$$
$$t = j + 2^{K-1}(i-1)\big\}, \quad t = 1,\ldots,n.$$

According to the truncated expansion (5), if only a limited number $M < N$ of eigenvectors are considered, the matrix $\hat{\mathbf{X}} = \sum_{i=1}^{M}\mathbf{X}_i$ provides the best approximation to the trajectory matrix $\mathbf{X}$, so that $\|\mathbf{X}-\hat{\mathbf{X}}\|$ is minimum and $\hat{v}(t) = \hat{\mu}(t) + \hat{x}(t)$ represents an estimate of the latent process $V(t)$.

Note that $\|\mathbf{X}\|^2 = \sum_{i=1}^{N}\hat{\lambda}_i$ and $\|\mathbf{X}_i\|^2 = \hat{\lambda}_i$, for $i = 1,\ldots,N$. Thus, we can consider the ratio, $\hat{\lambda}_i / \sum_{j=1}^{N}\hat{\lambda}_j$, as the characteristic of the contribution of matrix $\mathbf{X}_i$ to expansion (8). Accordingly, the sum of the first M ratios, $\sum_{i=1}^{M}\hat{\lambda}_i / \sum_{j=1}^{N}\hat{\lambda}_j$, is the characteristic of the optimal approximation of the trajectory matrix by the matrix of rank M.

2.3. The monthly energy consumption in Italy (1978–1995)

To show a simple example of eigentriple clustering, we consider the series of monthly energy consumption in Italy for the period 1978–1995. The time series shows a linear trend but a constant mean can be assumed for the residual series after removing

Table 1
Monthly energy consumption in Italy – Identification of temporal patterns and eigentriple aggregation

Frequencies	Period (months)	Harmonics components	Eigentriple subsets	Explained variance (%)	Test statistics
0.524	12	$S(t)$	1st–2nd	35.0	0.019
3.142	2	$S(t)$	3rd	17.9	–
2.094	3	$S(t)$	4th–5th	17.4	0.005
1.571	4	$S(t)$	7th–8th	12.2	0.006
1.047	6	$S(t)$	10th–11th	5.3	0.008
0.058	108	$C(t)$	6th–9th	4.5	0.001
2.618	2.5	$S(t)$	12th–13th	2.6	0.010
0.087	72	$C(t)$	14th	1.2	–

the trend by ordinary least squares. The residual series clearly shows the presence of harmonic components, and, in fact, both Whittle's and Hartley's tests (Priestley, 1981) confirm that eight periodogram ordinates are significantly large. By exploring the frequency content of the estimated principal components, we have then found that there are 14 eigentriples whose frequencies coincide with those of the residual series. The frequencies of the eight harmonics, the clustering of the eigentriples, their explained variance, and the values of the χ^2 test for each subset, g_k, are shown in Table 1. Note that taking $\alpha = 0.1$, for pairs of eigentriples (i.e., $r = 2$), we have two degrees of freedom and the critical value for the χ^2 test is 9.21.

3. Multiresolution Karhunen–Loéve

There are some cases of interest in which the components of a signal reside in nonoverlapping scales; in these cases, a multiresolution analysis is useful to highlight the latent features of the signal. The multiresolution Karhunen–Loéve (MR-KL) essentially computes the KL transform for successive levels of resolution. The MR-KL is applied similarly to a wavelet packet transform (WPT) (Mallat, 1998), in the sense that the KL transform is applied for each of the subsignals of the preceding level. The top level is the time representation of the signal. For ease of presentation, we summarize the hierarchical structure of the procedure in the following steps.

Step 1: For a signal $\mathbf{x}$ of length n, choose J and K to define the trajectory matrix $\mathbf{X}$ by equation (7). Since $rank(\widehat{\mathbf{R}}) \leq min(2^{J-K+1}, 2^K)$, it is reasonable to choose, $K \leq \left\lfloor \frac{J+1}{2} \right\rfloor$, to ensure that $2^{J-K+1} \geq 2^K$. Then, also define the maximum resolution level, $L \leq \frac{J-K}{K-1}$, so that for each level of resolution, l, the number of the *lagged vectors* is greater or at least equal to the dimension of the trajectory space.

Step 2: Compute the covariance matrix, $\widehat{\mathbf{R}}$, and obtain the $(2^{J-K+1} \times N)$ estimated principal component matrix, $\widetilde{\mathbf{Z}}^{(1)} = \mathbf{X}\mathbf{\Phi}$, where $\mathbf{\Phi} = \left(\widehat{\boldsymbol{\phi}}_1 \ldots \widehat{\boldsymbol{\phi}}_N\right)$ is the matrix of eigenvectors obtained by the eigen-decomposition of $\widehat{\mathbf{R}}$.

Step 3: For the resolution levels, $l = 2, 3, \ldots, L$, repeat the following steps:

(a) Set $\mathbf{z}_p^{(l)} = \tilde{\mathbf{z}}_p^{(l-1)}$, for $p = 1, \ldots, 2^{(l-1)K}$, where $\tilde{\mathbf{z}}_p^{(l-1)}$ denotes the pth column of $\widetilde{\mathbf{Z}}^{(l-1)}$;

(b) Following Eq. (7), obtain the $\mathbf{Z}_p^{(l)}$ matrix, with generic element given by

$$Z_p^{(l)}(i,j) = 2^{-1/2} z_p^{(l)}\left(j + 2^{K-1}(i-1)\right),$$

$$i = 1, \ldots, 2^{J-l(K-1)}, \quad j = 1, \ldots, N;$$

(c) Perform the spectral decomposition, $\hat{\mathbf{R}}_p^{(l)} = \mathbf{\Phi}_p^{(l)} \Lambda_p^{(l)} \mathbf{\Phi}_p^{(l)'}$, where $\hat{\mathbf{R}}_p^{(l)} = 2^{-J+l(K-1)} \mathbf{Z}_p^{(l)'} \mathbf{Z}_p^{(l)}$. Notice that at this stage, the following relationships hold for the eigenvalues:

$$\sum_{j=1}^{2^K} \hat{\lambda}_{p,j}^{(l)} = 2^{K-1} \hat{\lambda}_p^{(l-1)}, \quad \sum_{p=1}^{2^{(l-1)K}} \sum_{j=1}^{2^K} \hat{\lambda}_{p,j}^{(l)} = 2^{l(K-1)} \sigma_x^2;$$

(d) Obtain the $(2^{J-l(K-1)} \times 2^K)$ principal component matrix $\tilde{\mathbf{Z}}_p^{(l)} = \mathbf{Z}_p^{(l)} \mathbf{\Phi}_p^{(l)}$;

(e) Define the matrix of KL coefficients, $\widetilde{\mathbf{Z}}^{(l)} = \left[\tilde{\mathbf{Z}}_1^{(l)} \ldots \tilde{\mathbf{Z}}_p^{(l)} \ldots \tilde{\mathbf{Z}}_{2^{(l-1)K}}^{(l)}\right]$.

At each level, $l = 1, \ldots, L$, the signal $x_p^{(l)}(h)$ can be reconstructed as

$$x_p^{(l)}(h) = \sqrt{2}\, mean\left\{X_p^{(l)}(i,j) : j + 2^{K-1}(i-1) = h\right\},$$

$$i = 1, \ldots, 2^{J-l(K-1)}, \quad j = 1, \ldots, N;$$

where $X_p^{(l)}(i,j)$ is the (i,j)th element of $\mathbf{X}_p^{(l)} = \tilde{\mathbf{Z}}_p^{(l)} \mathbf{\Phi}_p^{(l)'}$. Furthermore, let $\tilde{\mathbf{z}}^{(l)} = vec\left(\widetilde{\mathbf{Z}}^{(l)}\right)$, then it can be shown (Fontanella et al., 2010) that at each level, l, the KL coefficients preserve the energy of the residual signal, that is, $\|\tilde{\mathbf{z}}^{(l)}\|^2 = \|\mathbf{x}\|^2$.

A summary scheme of MR-KL for $J = 6$, $L = 2$ and $K = 2$ is shown in Fig. 1. We notice that compared to the classical scheme of a discrete wavelet transform, where only the approximation space is decomposed, the full tree contains redundancy. This is not optimal for data compression but it is helpful to emphasize key structures in a signal. In fact, by decomposing the whole subsignal, $\mathbf{z}_p^{(l)}$, at the resolution level l, we can separate the frequency band uniformly and allow for a better frequency localization of the signal features. This also explains why the wavelet packet approach has been largely used to produce features suited to detection and discrimination (see, for example, Learned and Willsky (1995), Walczak et al. (1996), and references therein).

3.1. *Noise filtering*

Since all signals obtained as instrumental response of analytical apparatus are affected by noise, once the KL coefficients are available, a nonlinear approximation can be applied to recover the noise-free signal, $V(t)$ in Eq. (6) (Mallat, 1998). To allow for the splitting of the subspaces of signal and noise, all the KL coefficients defined at the highest resolution level, $\widetilde{\mathbf{Z}}^{(L)}$, are subject to a *thresholding* procedure based on their magnitude. We consider a *hard* procedure in which the threshold is defined as $\tau\sqrt{2^L}\hat{\sigma}_\epsilon$, where $\hat{\sigma}_e^2$ is the estimate of the noise variance provided at the first resolution level and τ is a suitable constant.

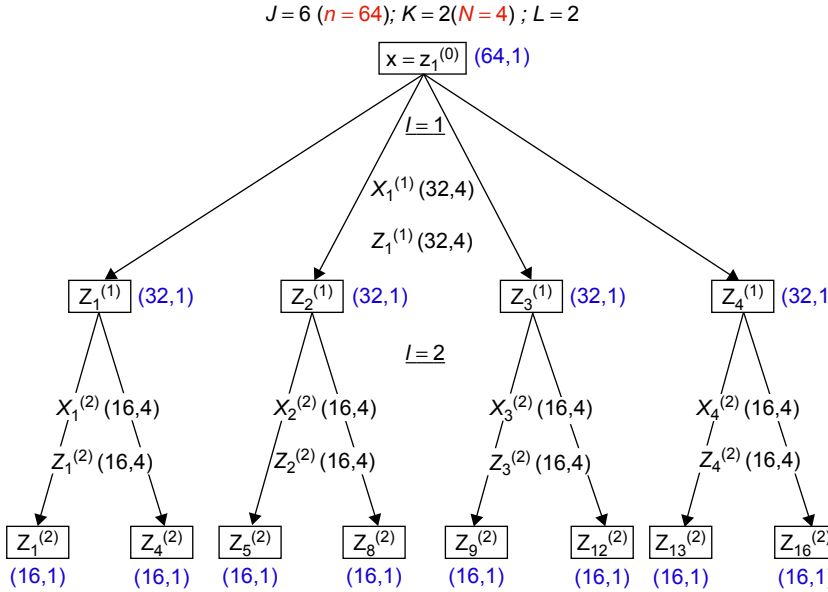

Fig. 1. Scheme of MR-KL for $J = 6$ and $K = 2$.

Different methods can be used to estimate σ_e^2, and for a discussion, see, for example, Ippoliti et al. (2005). One possibility is to estimate σ_e^2 as an average of the last r ordered eigenvalues, where r can be defined through Akaike's information theoretic criterion (AIC) as described by Fontanella et al. (2010). Then, since $V(t)$ is independent of $E(t)$, we may write $\hat{\sigma}_x^2 = \hat{\sigma}_v^2 + \hat{\sigma}_e^2$, and in terms of the principal components, it follows $\hat{\sigma}^2_{\tilde{\mathbf{z}}^{(L)}} = \hat{\sigma}^2_{\tilde{\mathbf{z}}_v^{(L)}} + \hat{\sigma}^2_{\tilde{\mathbf{z}}_e^{(L)}}$; then, since $\hat{\sigma}^2_{\tilde{\mathbf{z}}^{(L)}} = 2^{-L}\hat{\sigma}_x^2$, we also have $\hat{\sigma}^2_{\tilde{\mathbf{z}}_e^{(L)}} = 2^{-L}\hat{\sigma}_e^2$. If we assume, $\tilde{\mathbf{z}}_e^{(L)} \sim WN(0, 2^{-L}\sigma_e^2)$, then we have a probability equal to $1 - \alpha_\tau$ to observe values in the interval $\pm\tau\sqrt{2^L}\hat{\sigma}_e$. As a rule of thumb, simulation results discussed by Kostantinides and Yao (1988) show that the simple threshold, $\tau = 3$, performs more stably under a variety of noise levels. Values close to 3 for τ are also suggested in the study by Walker (1999, Section 2.6).

3.2. MR-KL Analysis of Infrared signals

The multiresolution Karhunen–Loéve analysis has been found useful in describing the dynamics of biomedical time series (see, for example, Fontanella et al. (2010)). In this section, we consider a psychophysiological study of the response of two subjects who underwent an emotional induction experiment. The sympathetic response to the external stimuli is analyzed through an infrared thermal (IR) signal, which gives a measure of the thermoregulatory actions in the forehead (for more information on thermal imaging see, for example, Shastri et al. (2009)). The experiment is described in the study by Fontanella et al. (2010), and we refer the interested reader to this work for full details. Suffice to say here that an audio and visual paradigm is used to elicit a startle response. A series of five different grey-tone images of human faces with a light grey background is used for visual stimulation. The inter-image interval lasts 18 s and each

image becomes visible for 2 s. The series of the face images is repetitively presented for a total of 11 cycles. The auditory stimulus is a 90 dB white noise burst lasting 200 ms, which is delivered along with the presentation of the fifth face of the series during the third, fourth, fifth, sixth, ninth, and tenth cycles. The thermal recording was done for the forehead region by means of a digital infrared camera. The objective of the analysis is the recognition of the features of the experiment by studying the sympathetic response of the two subjects. Because the IR signal depends on signal components, which in turn reside in nonoverlapping scales (Shastri et al., 2009), the multiresolution approach appears appropriate for this study.

The two series, each consisting of 1024 observations (one per second), show nonlinear trends, which are removed by smoothing splines. The original signals, the estimated trends, and the residual series are shown in Fig. 2.

For each subject, by setting $J = 10$ and $K = 3$, a (256×8) trajectory matrix, $\mathbf{X}$, is obtained from the residual signal. Then, following the procedure described above, the expansion coefficients are subject to a nonlinear thresholding to remove the measurement noise.

For both subjects, the first principal component, which in average accounts for almost 50% of the variability, appears quite smooth and reconstructs a signal that

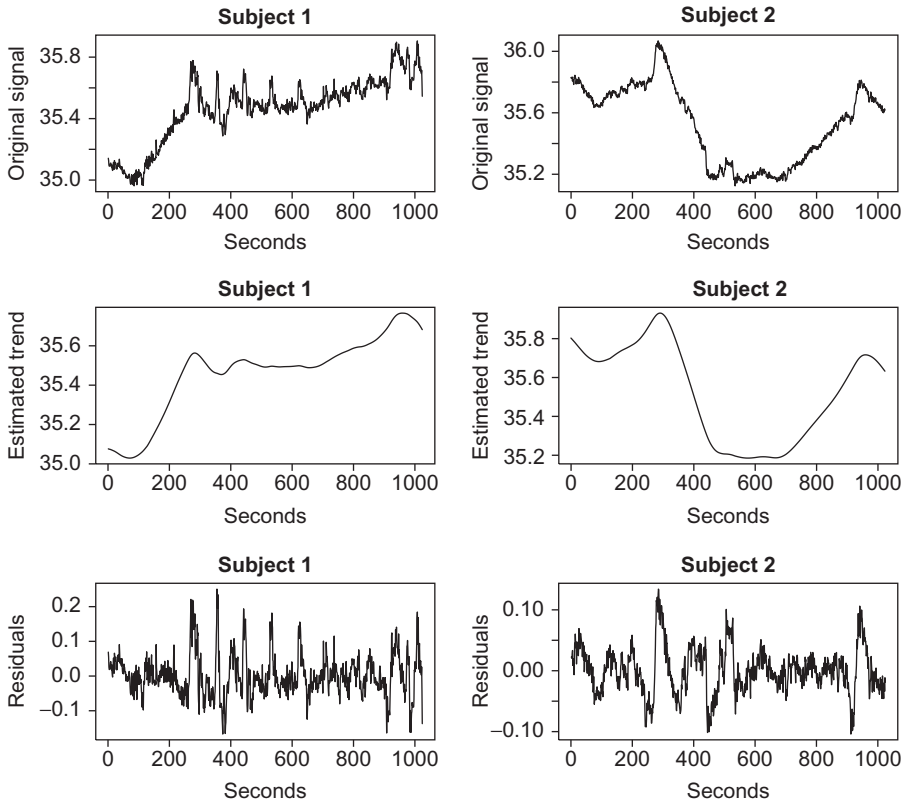

Fig. 2. Original IR series (top panels), estimated trends (centre panels), and residual series (bottom panels) for two subjects who underwent the emotional induction experiment.

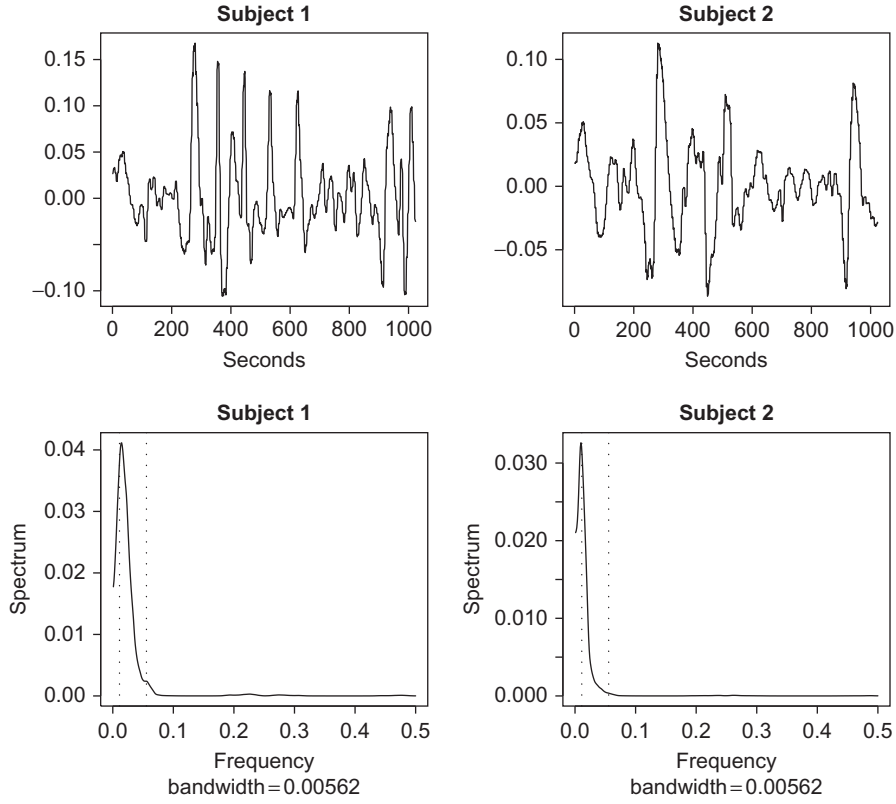

Fig. 3. Signals reconstructed by the first principal component (top panels) and the corresponding log-periodograms (bottom panels) smoothed by a Daniell window with truncation points (7, 7).

closely resembles the pattern of the residual series; also, it shows clear frequencies of interest. As a confirmatory analysis, the estimated spectra of the reconstructed signals shows the largest peak at a frequency corresponding to a period of 90 s, which in turn, represents the interstimuli interval within the third and sixth recurrences. Figure 3 shows the residual series reconstructed by the first principal component and the corresponding log-periodograms smoothed by a Daniell window with truncation points (7, 7).

The remaining principal components are much more "irregular" with a higher frequency content. Figure 4 shows the residual series reconstructed by the second principal component for which, in average, the explained variability is around 25%. We note that the corresponding log-periodograms, smoothed by a Daniell window with truncation points (7, 7), show the most significant peaks at the same period of the image sequences. Specifically, for both subjects, we have found that the estimated period is around 17.96 s, which is clearly associated to the inter-image interval. However, the possibility of detecting the effect of a specific image of the series appears difficult at this stage.

The residual signals reconstructed by the remaining components are characterized by even higher frequencies and smaller amplitudes. In average, we have observed that

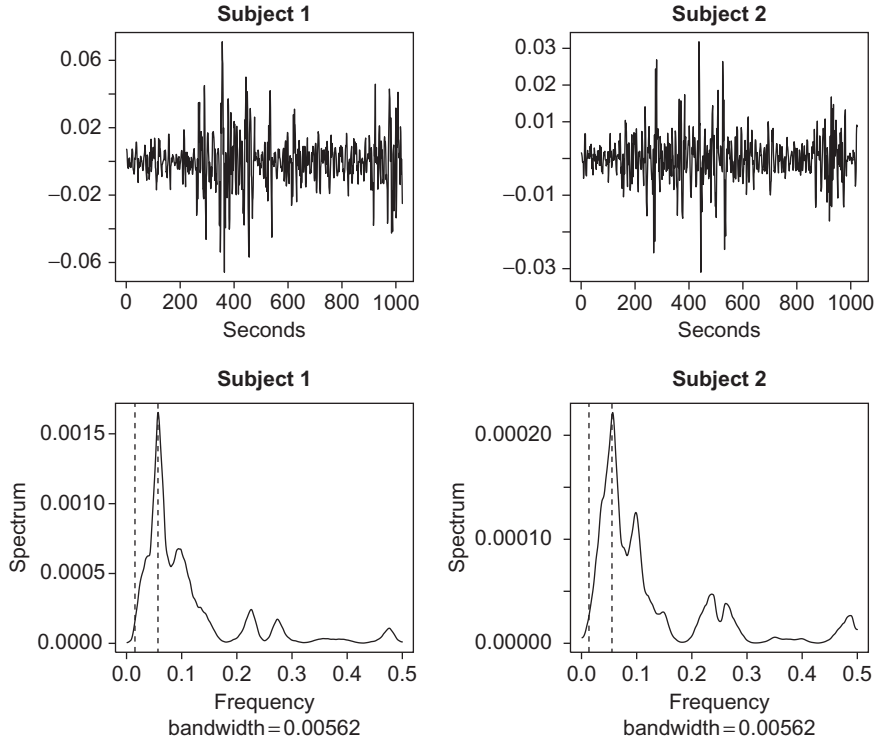

Fig. 4. Signals reconstructed by the second principal component (top panels) and the corresponding log-periodograms (bottom panels) smoothed by a Daniell window with truncation points (7, 7).

the first three components explain around 86% of the variability, and quite interestingly, the fifth and sixth components clearly carry on frequencies corresponding to a period of 3.3 s, presumably representing the breathing activity, which, as known, is characterized by 15–22 cycles per minute.

4. Karhunen–Loéve expansion of coupled one-dimensional processes

There are also cases of interest where coupled (correlated) stochastic processes are available. In these cases, it is possible to exploit the correlation between the two processes in the framework of simultaneous decomposition techniques, hence, extending the KL expansion to cases in which two kernels, R_1 and R_2, are defined. Let $R_1(t,t')$ and $R_2(t,t')$ denote two real, symmetric, and square integrable functions, and let R_1 and R_2 be the integral operators with kernels $R_1(t,t')$ and $R_2(t,t')$. Also assume that R_1 and R_2 are positive definite and non-negative definite, respectively, and that $R = R_1^{-1/2} R_2 R_1^{-1/2}$ is densely defined, bounded, and its extension to the whole of $L_2(T)$ has eigenfunctions that span $L_2(T)$. Then, if λ_i and ψ_i are the eigenvalues and the orthonormalized eigenfunctions of R, we have the following expansions (Kadota, 1967):

$$R_1(t,t') = \sum_i w_i(t) w_i(t') \quad \text{and} \quad R_2(t,t') = \sum_i \lambda_i w_i(t) w_i(t'),$$

where $w_i(t) = R_1^{1/2}\psi_i(t)$. The $w_i(t)$ also satisfies the following integral equation

$$\int_T R_2(t,t')w_i(t)dt = \lambda_i \int_T R_1(t,t')w_i(t)dt, \tag{9}$$

which represents an extension of the Fredholm integral (1). In fact, if R_1 and R_2 commute, then (9) reduces to (1).

In practice, for two matrices, $\mathbf{R}_1$ and $\mathbf{R}_2$, the simultaneous diagonalization of the two kernels can be approximated as

$$\sum_{i=1}^{M} R_2(t,t')w_i(t) = \lambda_i \sum_{i=1}^{M} R_1(t,t')w_i(t)$$

or in matrix formulation

$$\mathbf{R}_2\mathbf{w}_i = \lambda_i\mathbf{R}_1\mathbf{w}_i. \tag{10}$$

Equation (10) constitutes a *generalized eigenvalue decomposition* – GED – (Golub and Van Loan, 1993).

If $\mathbf{R}_2$ and $\mathbf{R}_1$ are symmetric and $\mathbf{R}_1$ is positive definite, then the eigenvalues λ_i and the eigenvectors $\mathbf{w}_i$ are real. Furthermore, if the eigenvalues are distinct, the different eigenvectors are orthogonal in the metrics $\mathbf{R}_2$ and $\mathbf{R}_1$

$$\mathbf{W}'\mathbf{R}_1\mathbf{W} = \mathbf{I}, \quad \text{and} \quad \mathbf{W}'\mathbf{R}_2\mathbf{W} = \mathbf{\Lambda},$$

where the columns of $\mathbf{W}$ consist of the eigenvectors $\mathbf{w}_i$ and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues, λ_i. If the matrix $\mathbf{R}_1$ is positive definite, Eq. (10) can be handled by the equivalent expression

$$\mathbf{R}_1^{-1}\mathbf{R}_2\mathbf{w}_i = \lambda_i\mathbf{w}_i.$$

In this case, the matrix $\mathbf{R}_1^{-1}\mathbf{R}_2$ is generally not symmetric, but it is possible to recover a symmetric eigenvalue problem using, for example, the Cholesky decomposition, $\mathbf{R}_1 = \mathbf{L}\mathbf{L}'$, and considering the eigenvalue decomposition of the symmetric matrix $\mathbf{R} = \mathbf{L}^{-1}\mathbf{R}_2(\mathbf{L}^{-1})'$. Its eigenvalues are the same of the original problem, whereas its eigenvectors are obtained as $\boldsymbol{\psi}_i = \mathbf{L}'\mathbf{w}_i$.

4.1. Types of kernels

In the following, we shall present three different criteria that emerge as solutions to special cases of the generalized eigenproblem (10). In all these cases, the following hold. Assume that $X_1(t)$ and $X_2(t)$ are two processes with mean zero and let $\mathbf{X}_1$ and $\mathbf{X}_2$ be the trajectory matrices obtained from the observed time series. Also denote with $\mathbf{R}_{x_1}$ and $\mathbf{R}_{x_2}$ the autocovariance matrices and with $\mathbf{R}_{x_1x_2}$ the cross-covariance matrix. Finally, let $\mathbf{w}_i = \left[\mathbf{w}'_{ix_1}\mathbf{w}'_{ix_2}\right]'$.

4.1.1. Partial Least Square

The goal of Partial Least Square (PLS) is to find the two directions of maximal data covariation (Naes and Martens, 1985); that is, the directions $\mathbf{w}_{x_i}$ and $\mathbf{w}_{y_i}$, such that the expansion coefficients, $\mathbf{z}_{ix_1} = \mathbf{X}_1\mathbf{w}_{ix_1}$ and $\mathbf{z}_{ix_2} = \mathbf{X}_2\mathbf{w}_{ix_2}$, have maximum covariance. Then, it can be shown (Fontanella et al., 2005) that the patterns $\mathbf{w}_i$ can be found through the generalized eigenvalue decomposition (10) with

$$\mathbf{R}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{R}_{x_1x_2} \\ \mathbf{R}_{x_2x_1} & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \mathbf{R}_1 = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

4.1.2. Canonical correlation analysis

The goal of canonical correlation analysis (CCA) is to find the two directions of maximal data correlation, that is, the directions $\mathbf{w}_{x_i}$ and $\mathbf{w}_{y_i}$, such that the expansion coefficients, $\mathbf{z}_{ix_1} = \mathbf{X}_1\mathbf{w}_{ix_1}$ and $\mathbf{z}_{ix_2} = \mathbf{X}_2\mathbf{w}_{ix_2}$, have the largest possible correlation (Mardia et al., 1979). Then, it can be shown (see Fontanella et al. (2005)) that the patterns $\mathbf{w}_{x_i}$ can be found through the generalized eigenvalue decomposition (10) with

$$\mathbf{R}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{R}_{x_1x_2} \\ \mathbf{R}_{x_2x_1} & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \mathbf{R}_1 = \begin{bmatrix} \mathbf{R}_{x_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{x_2} \end{bmatrix}.$$

4.1.3. Redundancy analysis

Given the two processes, X_1 and X_2, if the aim is to predict X_2 as well as possible in the least square error sense, the patterns $\mathbf{w}_{x_i}$ must be chosen, so that this error measure is minimized. This corresponds to a low-rank approximation of multivariate linear regression, which is also known as reduced rank regression (Izenman, 1975) or as redundancy analysis (RA) (van de Wollenberg, 1977).

Different from CCA and PLS, RA treats the two processes asymmetrically. In particular, RA seeks to find pairs of predictor and predictand patterns that maximize the predictand variance, and this is directly addressed by identifying patterns that are strongly related through the most efficient multivariate regression on X_2.

To measure the degree to which $\mathbf{X}_1$ can predict $\mathbf{X}_2$, the *redundancy index* can be used

$$R^2 = \frac{tr(\mathbf{R}_{\hat{x}_2})}{tr(\mathbf{R}_{x_2})} = \frac{tr(\mathbf{R}_{x_2x_1}\mathbf{R}_{x_1}^{-1}\mathbf{R}_{x_1x_2})}{tr(\mathbf{R}_{x_2})},$$

where tr denote the trace of the matrix. This index represents the proportion of the total variance in $\mathbf{X}_2$ that can be accounted for by the linear regression of $\mathbf{X}_2$ on $\mathbf{X}_1$.

In practice, it can be shown that the maximization of the redundancy index, and hence the identification of the best predicted and predictor patterns, is related to the solution of the generalized eigenvalue decomposition (10) with

$$\mathbf{R}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{R}_{x_1x_2} \\ \mathbf{R}_{x_2x_1} & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \mathbf{R}_1 = \begin{bmatrix} \mathbf{R}_{x_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

For a comparison of these techniques in atmospheric sciences, see, for example, Bretherton et al. (1992).

5. Karhunen–Loéve expansion of spatio-temporal processes

We conclude the chapter by showing the relevant theory of KL for spatio-temporal processes, which are continuous in space and discrete in time. Consider a spatio-temporal process $Y(\mathbf{s}_k, t)$, where $\mathbf{s}_k = \{s_{k1}, s_{k2}\} \in D$, with D some spatial domain in two dimensional Euclidean space $\Re^2$ and $t \in \{1, 2, \ldots, T\}$ a discrete index of times. At each time point, t, assume also that $X(\mathbf{s}_k, t) = Y(\mathbf{s}_k, t) - \mu(\mathbf{s}_k, t)$ is a zero-mean second-order spatial stochastic process with covariance function $R(\mathbf{s}_k, \mathbf{s}_j)$. Then, paralleling results shown in Section 2, $X(\mathbf{s}_k; t)$ can be expanded in any set of orthonormal basis functions, $\phi_i(\mathbf{s}_k)$, which are the eigenfunctions of the covariance function. Given a spatio-temporal process, KL analysis thus finds a set of orthogonal spatial patterns along with a set of associated uncorrelated time series. However, the difficulties of the approach are considerable for a continuous domain when data are collected only from a sparse and irregular network. The fact that we are considering a process observed at discrete points is a practical limitation to the numerical solution of (1). Accordingly, if there are p sample points in the domain, only p eigenfunctions can be estimated while, indeed, there are a denumerable infinity for a continuous process. Thus, the geometrical relations involving the domain of integration and the relations between the sites $\mathbf{s}_k, k = 1, \ldots, p$, are completely ignored in a discrete matrix formulation of (1). However, this limitation should be recognized as a restriction on the accuracy of the solution, but not as a part of the problem formulation. Hence, the numerical problem encountered in practice is to estimate $R(\mathbf{s}_k, \mathbf{s}_j)$ and attempt to solve Eqs (1)–(4). Obled and Creutin (1986) proposed a general approach based on a set of functions, $\{e_1(\mathbf{s}_i), e_2(\mathbf{s}_i), \ldots, e_p(\mathbf{s}_i)\}$, having a vector space structure over D. This approach leads to the following finite formulation of the Fredholm integral

$$\sum_{j=1}^{p} \sum_{m=1}^{p} R(\mathbf{s}_k, \mathbf{s}_j) E_{jm} \phi_i(\mathbf{s}_m) = \lambda_i \phi_i(\mathbf{s}_k), \quad i, k = 1 \ldots, p, \tag{11}$$

where $E_{jm} = \int_D e_j(\mathbf{s}) e_m(\mathbf{s}) d\mathbf{s}$, which is the *quadrature factor*. A finite solution of Eq. (4) is also

$$z_i(t) = \sum_{k=1}^{p} \sum_{j=1}^{p} X(\mathbf{s}_k, t) E_{kj} \phi_i(\mathbf{s}_j), \quad i = 1 \ldots, p. \tag{12}$$

The major difference between Eqs (11) and (1) is that in (11) we have to solve the problem of choosing a set of appropriate generating functions. From a practical point of view, the problem is limited to the evaluation of the integral of the E_{jm} term. In the two-dimensional case, Cohen and Jones (1969) and Buell (1972) suggested using piecewise constant functions. Following this approach, a set of areas of influence, $\{\delta(\mathbf{s}_k)\}$, $k = 1, \ldots, p$, for each site $\mathbf{s}_k$, is defined and, $e_m(\mathbf{s}_k)$, is assumed to be constant and equal to one over the area and zero elsewhere. One possibility is that the areas of influence are obtained by applying a Voronoi tessellation (Okabe et al., 1992) of D and each area can be taken to approximate the integral of the E_{jm} term. These areas compensate for the effects due to the variable density of the network. As a consequence, the numerical

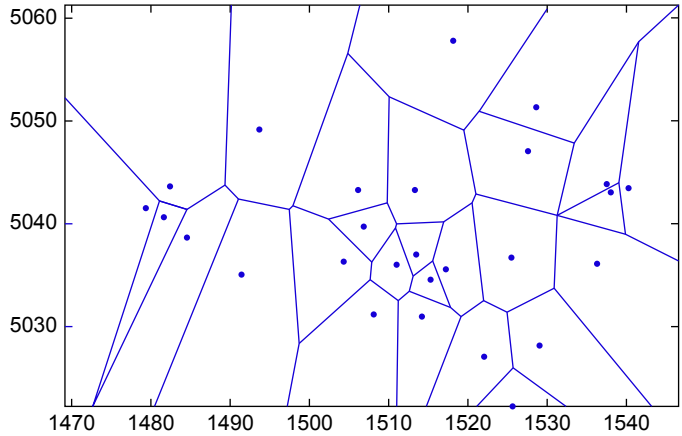

Fig. 5. Voronoi tessellation. The figure shows the areas of influence for the monitoring network used in the Milan district.

approximation of the Fredholm integral is

$$\sum_{j=1}^{p} R(\mathbf{s}_k, \mathbf{s}_j)\delta(\mathbf{s}_j)\phi_i(\mathbf{s}_j) = \lambda_i \phi_i(\mathbf{s}_k),$$

which can be rewritten in its symmetric form as

$$\sum_{j=1}^{p} R^*(\mathbf{s}_k, \mathbf{s}_j)\theta_i(\mathbf{s}_j) = \lambda_i \theta_i(\mathbf{s}_k),$$

where $\theta_i(\mathbf{s}_j) = \phi_i(\mathbf{s}_j)\sqrt{\delta(\mathbf{s}_j)}$ and $R^*(\mathbf{s}_k, \mathbf{s}_j) = R(\mathbf{s}_k, \mathbf{s}_j)\sqrt{\delta(\mathbf{s}_k)\delta(\mathbf{s}_j)}$.

As an example, Fig. 5 shows the Voronoi tessellation for some of the sites of the monitoring network used in the Milan district. The coordinate system is the Italian national grid system (Gauss–Boaga), which is based on the Universal Transverse Mercator (UTM) projection.

Note that when regular gridded fields are considered, then the quadrature factors are not needed and all follows as in Section 2.

5.1. *Computational details*

Assuming that the field is observed at p different sites and n temporal instants, and that the observed process can be represented by a $(n \times p)$ data matrix, $\mathbf{Y}$, von Storch and Zwiers (1999), Jolliffe (2002), and Wilks (2006) provide detailed descriptions of how to obtain EOFs through the singular value decomposition of the centred data matrix.

Here, we discuss a model-based approach that, essentially, represents a population approach to EOF. As in geostatistical analyses, we assume that the spatial covariance function is parameterized according to a valid spatial covariance function (Cressie, 1993). Assuming gaussianity, the spatial parameters can be estimated by

minimizing the deviance, minus twice the log-likelihood, over the valid parameter space

$$\mathcal{D}(\boldsymbol{\beta}) \propto -\frac{T}{2}\log|\mathbf{R}(\boldsymbol{\beta})| - \frac{1}{2}\sum_{t=1}^{T}\mathbf{x}(t)'\mathbf{R}(\boldsymbol{\beta})^{-1}\mathbf{x}(t).$$

where $\mathbf{R}(\boldsymbol{\beta})$ is the $(p \times p)$ spatial covariance matrix and, $\mathbf{x}(t) = \mathbf{y}(t) - \boldsymbol{\mu}(t)$, is the $(p \times 1)$ spatial series observed at time t. Note that estimation based on variogram functions is also possible and an example is described in the study by Sahu and Mardia (2005). Once the covariance function has been estimated, the eigen-decomposition of $\widehat{\mathbf{R}}(\boldsymbol{\beta})$ provides the set of singular values, $\hat{\lambda}_i$, and eigenvectors (spatial patterns), $\hat{\phi}_i(\mathbf{s}_k)$, $i = 1, \ldots, p$.

5.2. *State-Space formulation*

The linear Gaussian state-space model (Hamilton, 1994), combined with the KL (EOF) theory, provides a convenient way to produce spatial and spatio-temporal predictions of the field. The model we consider is described by the following *state* and *measurement* equations

$$\begin{aligned}\mathbf{z}(t) &= \Phi\,\mathbf{z}(t-1) + \boldsymbol{\epsilon}(t)\\ \mathbf{y}(t) &= \mathbf{H}\,\mathbf{z}(t) + \mathbf{u}(t),\end{aligned} \tag{13}$$

where $\mathbf{z}(t)$ is the *state* vector, Φ is the nonsingular *transition* matrix, $\mathbf{y}(t)$ is the measurement vector, and $\mathbf{H}$ is a constant output matrix. The sequences, $\boldsymbol{\epsilon}(t)$ and $\mathbf{u}(t)$, are assumed to be mutually independent, normally distributed random variables and represent the state and measurement errors, respectively.

To provide an example of model specification, assume for simplicity that, at each time t, the process shows a spatial linear trend. Also, assume that $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p$ and that the corresponding eigenvectors, $\hat{\phi}_i(\mathbf{s}_k)$, are sorted accordingly. Then, using the first $M \ll p$ eigenvectors, the measurement matrix, $\mathbf{H}$, appears as follows

$$\mathbf{H} = \begin{bmatrix} 1 & s_{11} & s_{12} & \hat{\phi}_1(\mathbf{s}_1) & \hat{\phi}_2(\mathbf{s}_1) & \cdots & \hat{\phi}_M(\mathbf{s}_1) \\ 1 & s_{21} & s_{22} & \hat{\phi}_1(\mathbf{s}_2) & \hat{\phi}_2(\mathbf{s}_2) & \cdots & \hat{\phi}_M(\mathbf{s}_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & s_{p1} & s_{p2} & \hat{\phi}_1(\mathbf{s}_p) & \hat{\phi}_2(\mathbf{s}_p) & \cdots & \hat{\phi}_M(\mathbf{s}_p) \end{bmatrix}$$

and the measurement Eq. (14) thus represents a truncated expansion as in (4). Note that the first three columns of $\mathbf{H}$ specify the regressors for the trend. Following Mardia et al. (1998), the columns of $\mathbf{h}_k$, $\mathbf{H}$, are known as *common fields*, and since they have a spatial structure, they can be regarded as spatial series. To provide an example, assume that the field is observed on a (16×16) regular lattice. Assume also that the spatial covariance function is "spherical" (Cressie, 1993) with parameters: *range* 10, *partial sill* 5, and *nugget* 0.1. Then, the eigendecomposition of the spatial covariance matrix provides a

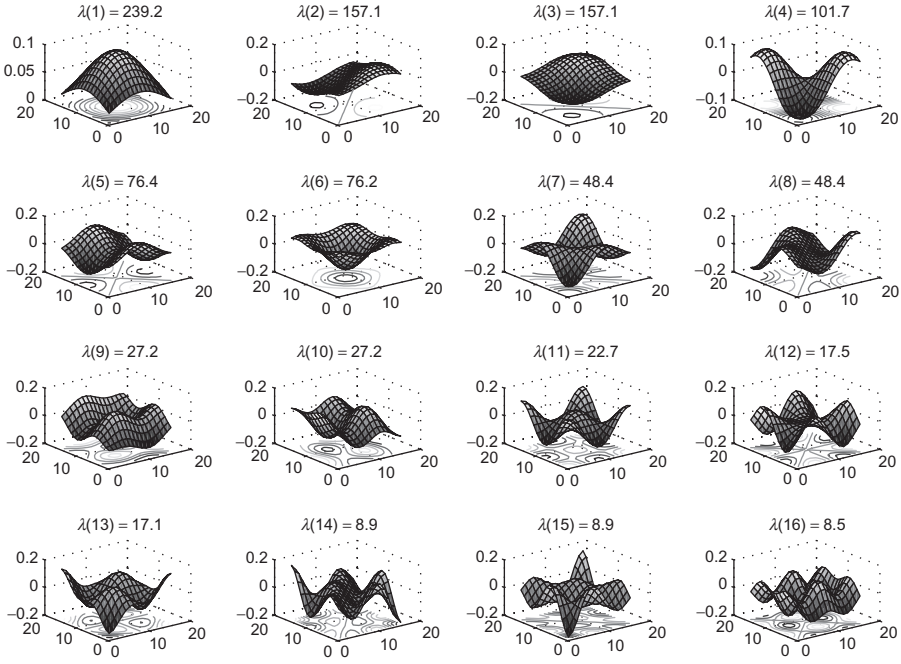

Fig. 6. The spatial patterns of the first 16 eigenvectors obtained from the eigendecomposition of a "spherical" covariance matrix. The parameters of the covariance function are *range* 10, *partial sill* 5, and *nugget* 0.1.

series of 256 eigenvectors that are ordered according to the magnitude of the corresponding eigenvalues. Figure 6 shows the spatial patterns of the first 16 eigenvectors. The figure suggests that the first eigenvectors, corresponding to the largest eigenvalues, are smooth and show a low frequency content; they are thus able to capture the large-scale variation of the process. On the other hand, we note that as the eigenvalues decrease, the spatial patterns of the corresponding eigenvectors become much more irregular. Thus, the last eigenvectors provide information on the small-scale variation of the field. Thus, M acts as a regularization parameter and for M being sufficiently small, only a few number of expansion coefficients, $\mathbf{z}(t)$, have to be estimated through the recursion of the Kalman filter (Hamilton, 1994).

This model specification, initially used in the study by Fontanella and Ippoliti (2003), represents a simple version of the models discussed by Mardia et al. (1998), Wikle and Cressie (1999), and Sahu and Mardia (2005) with applications in environmental sciences.

5.2.1. *Spatial and spatio-temporal predictions*

For $t \leq T$, the state-space formulation can be used to fit the data or, eventually, to reconstruct missing data. Interpolation at an unobserved spatial location, $\mathbf{s}_0$, is also possible. A straightforward approach is to use the following equation

$$\hat{y}(\mathbf{s}_0, t) = \sum_{k=1}^{3+M} h_k(\mathbf{s}_0) z_k(t), \tag{14}$$

which essentially requires the interpolation of the M spatial patterns, $\boldsymbol{\phi}_i(\mathbf{s}_0)$, at site $\mathbf{s}_0$. This prediction is not a difficult task, and, ensuring orthogonality, we could apply relatively simple interpolation schemes, such as thin-plate splines. Mardia and colleagues (1998) and Wikle and Cressie (1999) discuss two alternative approaches.

Temporal predictions of Y are ensured by the dynamic of the state equation; in fact, k-step ahead forecasts of the expansion coefficients, $\mathbf{z}(t)$, can be obtained as

$$\hat{\mathbf{z}}(t+k|T) = \Phi^k \, \mathbf{z}(t), \quad t \geq T$$

and forecasts of Y are obtained by plugging in $\hat{\mathbf{z}}(t+k|T)$ in the measurement equation

$$\hat{\mathbf{y}}(t+k|T) = \mathbf{H} \, \hat{\mathbf{z}}(t+k|T) \tag{15}$$

with prediction variance

$$Var\Big(\mathbf{z}(t) - \hat{\mathbf{z}}(t+k|T)\Big) = \mathbf{H}(\Phi^k \mathbf{P}_{T|T} \Phi^{k'} + \Sigma_\epsilon)\mathbf{H}' + \Sigma_\mathrm{u},$$

where Σ_ϵ and Σ_u are the covariance matrices of $\boldsymbol{\epsilon}(t)$ and $\mathbf{u}(t)$, respectively, and $\mathbf{P}_{T|T}$ is the variance prediction error (computed by the Kalman filter) of the state vector. By combining Eqs (14) and (15), we can finally obtain spatio-temporal predictions.

6. Discussion

This chapter has illustrated that the KL technique performs well in the extraction of specific features of temporal and spatio-temporal data. We have also shown that KL (and EOF) analysis is a useful tool for dimensionality reduction. We began by reviewing the conventional KL method for one-dimensional processes; then, we described the *decomposition* and *reconstruction* phases to illustrate the specific steps of the analysis. Of course, a common goal of time series analysis is extrapolating past behavior into the future. Here, we have not considered the forecasting problem but specific details, including how to specify forecast confidence bounds, are given in the study by Golyandina et al. (2001, Section 2.4).

An important part of data modeling is the specification of the trajectory matrix and its parameter K, which defines the window length and, hence, the number N of the delayed copies of the series. The numerical value of K is determined experimentally, because in practice, its choice is guided by both the length of the signal and the number of components thought to be present in $V(t)$. The choice of K is also very much like to the choice of the length of the support of the wavelets, for example in a Daubechies wavelet transform.

A multiresolution version of the KL was also discussed. MR-KL allows for a nonlinear approximation, which is better suited for denoising purposes. The link between the MR-KL and other well-known multiresolution decompositions, including wavelets, may be examined by employing the system approach proposed by Unser (1993). However, MR-KL is characterized by basis functions, which are data adaptive. In contrast with wavelets and Fourier analysis, the KL model does not require an advance specification of the functional form of the eigenfunctions leaving it to be freely determined

by the structure of the data. As shown in Section 4, the possibility of deriving the basis functions from the covariance structure of the data allowed us to extend the theory within the generalized eigendecomposition. For example, this is particularly useful when two signals are available as described, for example, in the study by Merla et al. (2004) and Shastri et al. (2009).

The chapter also discussed the application of KLE in a spatio-temporal context. The expansion of the process has been defined within a state-space framework that allows to estimate the expansion coefficients through the Kalman recursions. Note that this approach contrasts with that described by Hannachi et al. (2007), where the spatial patterns and the expansion coefficients are obtained through the singular value decomposition of the spatio-temporal matrix.

Obviously, it is difficult to provide a total overview of a field that is too broad for us to be exhaustive. For example, we have not discussed some other extensions of EOFs including cyclostationary, PXEOFs, the S-mode EOF analysis, trend EOFs, and nonlinear extensions of PCA. Also, we have not used this chapter to fully describe the use of KL (EOF) in atmospheric sciences. For all these points, there are several reference books and review papers and we refer the interested reader to them for specific details.

Acknowledgments

The authors thank Dr. A. Merla from the Institute of Advanced Biomedical Technologies, University G. d'Annunzio (Italy), for providing the Infrared signals analysed in Section 3.2.

References

Anderson, T.W., 1963. Asymptotic theory for principal components analysis. Ann. Math. Stat. 34, 122–148.

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC, Boca Raton, FL.

Basilevsky, A., Hum, D.P.J., 1979. Karhunen-loeve analysis of historical time series with an application to plantation Birthsin Jamaica. J. Am. Stat. Assoc. 74, 284–290.

Box, G.E.P., Jenkins, G.H., 1970. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Bretherton, C.S., Smith, C., Wallace, J.M., 1992. An intercomparison of methods for finding coupled patterns in climate data. J. Clim. 5, 541–560.

Buell, C.E., 1972. Integral equation representation for factor analysis. J. Atmosph. Sci. 28, 1502–1505.

Cohen, A., Jones, R.H., 1969. Regression on a random field. JASA 64, 1172–1182.

Coli, M., Fontanella, L., Granturco, M.G., 2005. Parametric estimation for ARFIMA models via spectral methods. Stat. Meth. Appl. 14, 11–27.

Craddock, J.M., 1973. Problems and prospects for eigenvector analysis in meteorology. Statistician 22, 133–145.

Cressie, N., 1993. Statistics for Spatial Data. Wiley, New York.

Fontanella, L., Ippoliti, L., 2003. Dynamic models for space-time prediction via Karhunen-Loéve expansion. Stat. Meth. Appl. 12, 61–78.

Fontanella, L., Ippoliti, L., Mardia, K., 2005. Exploring apatio-temporal variability by eigen-decomposition techniques. In: Proceedings of the Meeting of the Italian Statistical Society: Statistics and Environment, 21–23 September 2005, Messina, Italy.

Fontanella, L., Ippoliti, L., Merla, A., 2010. Multiresolution Karhunen Loéve analysis of Galvanic skin response for psycho-physiological studies. Metrika (Online First). doi:10.1007/s00184-010-0327-3.

Ghanem, R., Spanos, P.D., 1991. Stochastic Finite Element: A Spectral Approach. Springer, New York.

Golub, G.H., Van Loan, C.F., 1993. Matrix Computations. John Hopkins University Press, Baltimore, Maryland.

Golyandina, N., Nekrutkin, V., Zhigljavsky, A., 2001. Analysis of Time Series Structure: SSA and Related Techniques. Chapman & Hall/CRC, New York/London.

Grenander, U., 1976. Pattern Synthesis: Lectures in Pattern Theory 1. Springer-Verlag, New York.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton.

Hannachi, A., Jolliffe, I.T., Stephenson, D.B., 2007. Empirical orthogonal functions and related techniques in atmospheric science: A review. Int. J. Climatol. 27, 1119–1152.

Ippoliti, L., Romagnoli, L., Fontanella, L., 2005. A Noise estimation method for corrupted correlated data. Stat. Meth. Appl. 14, 343–356.

Izenman, A.J., 1975. Reduced-rank regression for the multivariagte linear model. J. Multivar. Anal. 5, 248–264.

Jolliffe, I.T., 2002. Principal Component Analysis, second ed. Springer. New York.

Kadota, T.T., 1967. Simultaneous diagonalization of two covariance kernels and application to second order stochastic processes. SIAM J. Appl. Math. 15, 1470–1480.

Karhunen, K., 1947. Uber linear methoden in der wahrscheinlichkeitsrechnung. Ann. Acad. Sci. Fenn. (AI) 37, 1–79.

Kostantinides, K., Yao, K., 1988. Statistical analysis of effective singular values in matrix rank determination. IEEE Trans. Acoust. Speech Signal Process. 36, 757–763.

Learned, R.E., Willsky, A.S., 1995. A wavelet packet approach to transient signal classification. Appl. Comput. Harmon. Anal. 2, 256–278.

Li, C.C., Der-Kiureghian, A., 1993. Optimal discretization of random processes. J. Eng. Mech. 119, 1136–1154.

Loève, M., 1978. Probability Theory, vol. 2, fourth ed. Springer Verlag, New York.

Lorenz, E.N., 1956. Empirical Orthogonal Functions and Statistical Weather Prediction, Technical report, Statistical Forecast Project Report 1, Dep. of Meteor, MIT: 49.

Mallat, S., 1998. A Wavelet Tour of Signal Processing. Academic Press, New York.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, London.

Mardia, K.V., Redfern, E., Goodall, C.R., Alonso, F., 1998. The Kriged Kalman filter. TEST 7, 217–285.

Merla, A., Di Donato, L., Rossini, P.M., Romani, G.L., 2004. Emotion detection through Functional Infrared Imaging: preliminary results. Biomed. Tech. 48, 284–286.

Monahan, A.H., Fyfe, J.C., Ambaum, M.H.P., Stephenson, D.B., North, G.R., 2009. Empirical orthogonal functions: the medium is the message. J. Clim. 22, 6501–6514.

Naes, T., Martens, H., 1985. Comparison of prediction methods for multicollinear data. Commun. Stat. Simul. Comput. 14, 545–576.

Obled, C., Creutin, J.D., 1986. Some developments in the use of empirical orthogonal functions for mapping meteorological fields. J. Clim. Appl Meteorol. 25, 1189–1204.

Okabe, A., Boots, B., Sugihara, K., 1992. Spatial Tessellations. Concepts and Applications of Voronoi Diagrams. John Wiley & Sons. Chichester.

Priestley, M.B., 1981. Spectral Analysis and Time Series. Academic Press, London.

Riesz, F., Sz-Nagy, B., 1955. Functional Analysis. Ungar, New York.

Sahu, S.K., Mardia, K.V., 2005. A Bayesian Kriged-Kalman model for short-term forecasting of air pollution levels. J. Royal Stat. Soc. Ser. C 54, 223–244.

Shastri, D., Merla, A., Tsiamyrtzis, P., Pavlidis, I., 2009. Imaging facial signs of neurophysiological responses. IEEE Trans. Biomed. Eng. 56, 477–484.

Sherman, M., 2011. Spatial Statistics and Spatio-Temporal Data. Covariance Functions and Directional Properties. John Wiley & Sons, Chichester.

Shorack, G.R., Wellner, J.A., 1986. Empirical Processes with Applications to Statistics. Wiley, New York.

Unser, M., 1993. An extension of the Karhunen-Loéve transform for wavelets and perfect reconstruction filterbanks. Math Imaging SPIE 2034, 45–56.

van de Wollenberg, A.L., 1977. Redundancy analysis: an alternative for canonical correlation analysis. Psycometrica 36, 207–209.

von Storch, H., Zwiers, F.W., 1999. Statistical Analysis in Climate Research. Cambridge University Press, Cambridge.

Walczak, B., van den Bogaert, B., Massart, D.L., 1996. Application of wavelet packet transform in pattern recognition of near-IR data. Anal. Chem. 68, 1742–1747.

Walker, J.S., 1999. A Primer on Wavelets and their Scientific Applications.Chapman and Hall, CRC, Boca Raton.
Wikle, C.K., Cressie, N., 1999. A dimension-reduction approach to space-time Kalman filtering. Biometrika 86, 815–829.
Wilks, D.S., 2006. Statistical Methods in the Atmospheric Sciences, second ed. Academic Press, Amsterdam.
Zhang, J., Ellingwood, B., 1994. Orthogonal series expansions of random processes in reliability analysis. J. Eng. Mech. 120, 2660–2677.

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00018-1

Statistical Analysis of Spatio-Temporal Models and Their Applications

T. Subba Rao[1,2] *and Gy. Terdik*[3]

[1]*School of Mathematics, The University of Manchester, Manchester M13 9Pl, UK*
[2]*C. R. RAO AIMSCS, University of Hyderabad campus, Hyderabad, India*
[3]*Faculty of Informatics, University of Debrecen, 4010 Debrecen, Pf. 12, Hungary*

Abstract

In this chapter, we briefly review existing literature on Kriging of spatial random processes. In order to define a nonlinear type of Kriging estimator, we introduce measures of nonlinear dependence from the point of view of Kriging. Although we propose a methodology for testing, the distribution theory of the test statistics need to be investigated. We consider spatio-temporal processes at several locations and defining discrete Fourier transforms taken over the time series data at each location, we define simultaneous autoregressive spatio-temporal autoregressive (SAST) models and conditional spatio-temporal autoregressive models (CAST) in terms of these complex-valued random processes. These are similar to simultaneous autoregressive models of Whittle (Whittle, P., 1954. On stationary processes in the plane. Biometrika 49, 305–314) and conditional autoregressive models considered by Bartlett (Bartlett, M.S., 1978. Nearest neighbour models in the analysis of field experiments. J. R. Stat. Soc. Ser. B 40, 147–174) and Besag (Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B 36, 192–225). We outline an approach for the estimation of the models. We describe recent results by the authors and their co-authors on Space–time autoregressive models.

Keywords: Spatial, spatio-temporal processes, linear and nonlinear Kriging predictors, discrete Fourier transforms, SAST and CAST models, space time autoregressive models.

1. Introduction and basic ideas

Let $\{Z(s),\, s \in D \subset \mathbb{R}^d\}$ be a real-valued random process, where D is a fixed domain and an open set. Let $\{Z(s_i);\, i = 1, 2, \ldots, n\}$ be a sample from the random process, where the locations $(s_1, s_2, \ldots, s_n)$ are fixed. The process $\{Z(s)\}$ is also defined as a random field. Let us assume that the random process is spatially covariance stationary. Further conditions of spatial stationarity for defining higher order moments will be needed when we consider nonlinear Kriging. We say the process $\{Z(s)\}$ is second-order covariance stationary if

1. $E(Z(s_i)) = \mu$, for all i,
2. $Var(Z(s_i) = E(Z(s_i) - \mu)^2 < \infty$, for all i,
3. $Cov(Z(s_i), Z(s_j)) = C(s_i - s_j)$, is a function of the lag difference of the locations.

The covariance function $C(s_i - s_j)$ is non-negative definite. If $C(s_i - s_j) = R\left(\|s_i - s_j\|\right)$, where $\|s_i - s_j\|$ is the Euclidean distance, the spatial process is said to be isotropic. For the process to be isotropic, the process must be second-order stationary. Let us define the set $N(h) = \left\{(s_i, s_j);\, s_i - s_j = h\right\}$. In other words, the set $N(h)$ contains all the pairs (s_i, s_j) such that $s_i - s_j = h$. Let $\|N(h)\|$ be the total number of such pairs in the set, i.e., its cardinality. If μ, σ^2, and $C(s_i - s_j)$ are unknown, they can be estimated by

$$\hat{\mu} = \overline{Z} = \frac{1}{n}\sum Z(s_i), \quad \widehat{\sigma}^2 = \frac{1}{n}\sum (Z(s_i) - \overline{Z})^2$$

and the covariance function of lag difference "h" is estimated by

$$\widehat{C}(h) = \frac{1}{\|N(h)\|}\sum_i \sum_j (Z(s_i) - \overline{Z})(Z(s_j) - \overline{Z}),$$

where the summation is taken overall pairs defined in the set $N(h)$. The sampling properties of the estimators were discussed by Cressie (1993) (see references therein). In real spatial data analysis, the autocovariances $C(h)$ of the process play a role similar to sample autocovariances in time series analysis, for example, for model identification, diagnostic checking, etc. However, in contrast to time series, another important function that is often used in spatial analysis is "variogram" (or semi-variogram). The semi-variogram is defined as follows,

$$\gamma(s_i, s_j) = \frac{1}{2}E\left[Z(s_i) - Z(s_j)\right]^2.$$

Here $2\gamma(s_i, s_j)$ is known as the "variogram", and it is known that it must be conditionally non-negative definite, see Cressie (1993). From now onward, for convenience, we assume the mean $\mu = 0$. If $\gamma(s_i, s_j) = \gamma(s_i - s_j)$, then the random process $\{Z(s)\}$ is said to be intrinsically stationary. It is widely believed in spatial literature that the assumption that the process is stationary is unrealistic, but differenced process is stationary. If the random process is stationary, then

$$\gamma(s_i, s_j) = \gamma(s_i - s_j) = C(0) - C(h),$$

where $C(h) = C(s_i - s_j)$. If the covariance $C(h)$, a function of the distance "h", is a function of the Euclidean distance $\|h\|$, then we say the process is said to be isotropic. This isotropic covariance function is denoted by $R(\|h\|)$. If it is not isotropic, the process is said to be anisotropic. One of the classical differences between classical time series and spatial processes is that in time series (defined on the real line) one can define directionality, i.e., past, present, and future and this is not that obvious in general space. This lack of directionality is a serious problem and a stumbling block in modeling spatial processes.

Briefly we summarize the topics we considered in the following sections. In Section 1.1, we briefly outline various methods of estimation (prediction) of an observation at a known location. These predictors also called Kriging predictors. In view of the fact that these predictors may not be optimal in non-Gaussian situations, in Section 1.4 we have introduced quadratic predictors. We also pointed out that the performance of these new quadratic predictors need to be investigated. The sampling distribution of the test statistics for testing hypothesis need also investigation. We define measures of nonlinear dependence in Section 2. A brief outline of the frequency domain approach to random processes on lattices is given in Section 4, and the models defined in this section using discrete Fourier transforms are extended to include temporal dimension as well, and these are considered in Section 5. In Section 6, space–time linear ARMA models and their extension to nonlinear situations, leading to bilinear space–time models, are briefly discussed in Section 6.

1.1. Linear kriging (linear simple Kriging predictor)

One of the important objects of spatial process is the estimation of $Z(s_0)$, where the location s_0 is known, given a sample $\{Z(s_i); i = 1, 2, \ldots, n\}$ from $\{Z(s)\}$. Let $\mathbf{Z}'(\mathbf{s}) = (Z(s_1), Z(s_2), \ldots, Z(s_n))$, $\sigma'(s_0, \mathbf{s}) = (E(Z(s_0)Z(s_1)), E((Z(s_0)Z(s_2)), \ldots, E((Z(s_0)Z(s_n))$, and further let us define a matrix of order $n \times n$,

$$\mathbf{C} = \left(C\left(s_i, s_j\right)\right) = \left(C\left(s_i - s_j\right)\right).$$

In view of our assumption of spatial stationarity, each element of the matrix is a function of the spatial difference only, not location dependent. Let the estimate of $Z(s_0)$ be a linear combination of all the elements of $\mathbf{Z}(\mathbf{s})$, i.e., $\widehat{Z}(s_0) = \beta'\mathbf{Z}(\mathbf{s})$. The object is to find the vector β such that the mean square error is minimum.

Let us minimize

$$Q(\beta) = E\left(Z(s_0) - \beta'\mathbf{Z}(\mathbf{s})\right)^2 \text{ with respect to } \beta.$$

It can easily be shown that β must satisfy the equation $\mathbf{C}\beta = \sigma(s_0, \mathbf{s})$ and, if we assume that $\mathbf{C}$ is nonsingular, then we obtain $\beta = \mathbf{C}^{-1}\sigma(s_0, \mathbf{s})$. Hence, the linear estimate of $Z(s_0)$ is

$$\widehat{Z}(s_0) = \sigma'(s_0, \mathbf{s})\mathbf{C}^{-1}\mathbf{Z}(\mathbf{s}),$$

and the minimum is $\text{Min}\, Q(\beta) = \sigma^2 - \sigma'(s_0, \mathbf{s})\mathbf{C}^{-1}\sigma(s_0, \mathbf{s})$. We shall denote this minimum by Q_{lin}. In order to estimate $Z(s_0)$, we need estimators of σ^2, $\sigma(s_0, \mathbf{s})$, and $\mathbf{C}$.

The usual way is to define a parametric function for the spatial covariance, and estimate its parameters and use this estimated function to estimate $Z(s_0)$. We believe that, alternatively, one can estimate the elements of $\mathbf{C}$ and $\sigma(s_0, \mathbf{s})$ directly from the data as the elements of this vector depend on the difference of spatial locations rather than locations itself. Thus, we do not need observations at the location s_0. In view of this, we can consider those covariances with these differences as the elements of $\sigma(s_0, \mathbf{s})$. In other words, the lag or the Euclidean distances are important rather than locations. We note that the above predictor is linear, and if the process is Gaussian it is optimal. It is well known that the linear predictors are optimal only in Gaussian case. In other non-Gaussian cases, one has to consider nonlinear predictors to see whether we get better Kriging estimator. We mention here that the kriging estimator can be expressed in terms of valid variogram as well; for asymptotic sampling properties of this predictor and related references, see Lahiri et al. (2002).

1.2. Linear ordinary kriging estimator

In the above derivation, we assumed that the mean $\mu = 0$. If $\mu \neq 0$, then the modified predictor would be,

$$\widehat{Z}(s_0) = \mu + \sigma'(s_0, \mathbf{s})\mathbf{C}^{-1}(\mathbf{Z}(\mathbf{s}) - \mu\mathbf{1}),$$

where $\mathbf{1}' = (1, 1, \ldots .1)$, and the minimum mean square error remains the same since the mean is assumed to be known, i.e., Q_{lin}. The above predictor is usually known as the ordinary Kriging estimator.

1.3. Linear universal kriging estimator

A more realistic situation is to assume that the mean is a function of the location, i.e., $EZ(s) = \mu(s)$, possibly to accommodate the trend. A specific function that is often used is a polynomial (for $d = 2$)

$$\mu(s_i) = \sum\sum \alpha_{ll'} x_i^l y_i^{l'}, \quad l + l' \leq p, \quad (i = 1, 2, \ldots, n),$$

and in this case $\mu(s)$ is dependent on location coordinates $s_i = (x_i, y_i)$ (Cartesian coordinates) in terms of polynomial up to order p. In other words $\mu(s_i) = \mathbf{x}(s_i)\,\alpha$, where $\mathbf{x}(s_i)$ is a set of explanatory variables. The coefficients α are same for all locations. Of course, the regression parameters $\alpha = \{\alpha_{lm}\}$, need to be estimated as well. With this choice of the mean, the universal Kriging estimator will be of the form

$$\widehat{Z}(s_0) = \mu(s_0) + \sigma'(s_0, \mathbf{s})\mathbf{C}^{-1}(\mathbf{Z}(\mathbf{s}) - \boldsymbol{\mu}(\mathbf{s}))$$

As before, the minimum mean square error remains the same, provided no estimation is done.

1.4. General comments

As pointed out earlier, the above estimators depend on the knowledge of $\sigma(s_0, \underline{\mathbf{s}})$ and $\mathbf{C}$. Since we do not have any observations at s_0, the estimation of $\sigma(s_0, \mathbf{s})$ does not seem to be possible. But we believe that since elements of the vector depends only on distances, we could, as a first approximation, replace these elements by the sample covariances of the same lag differences. However, the usual practice is to assume a parametric form for the covariances (or variograms) and use these for the estimation purposes. These parametric forms depend on some unknown parameters, such as range parameters, smoothness parameters, etc. One of the important problems that is receiving considerable attention recently is finding the best set of methods for the estimation of these parameters. Briefly, we describe some approaches recently advocated for the estimation. By choosing an appropriate function for the variogram, we are reducing the parameters to be estimated. However, this leads to an important problem, namely, how to choose the best parametric function? One way to choose is to use cross validation methodology (see Cressie (1993) and Das (2011)). Several of the functions proposed in the literature belong to Matern (1986) class, which are briefly described below. Let $R(\|h\|)$ denote the covariance between random processes defined at two locations with lag difference "$h \in \mathbb{R}^d$". The Matern class of covariance function is

$$R(\|h\|) = \sigma^2(\Gamma(\nu))^{-1}\left(\frac{\theta\,\|h\|}{2}\right)^{\nu} 2K_\nu(\theta\,\|h\|),$$

$\nu > 0$, $\theta > 0$, where $K_\nu(\theta\,\|h\|)$ is a modified Bessel function of second order, and θ govern the range of spatial dependence. Here the parameter ν is the smoothness parameter governing the smoothness of the random process. To get some preliminary idea of the range parameter, usually one inspects the sample semi-variogram

$$\hat{\gamma}(h) = \frac{1}{\|N(h)\|}\sum_i (Z(s_i + h) - Z(s_i))^2;$$

and is based on the observation that as $\|h\| \to \infty$, $R(\|h\|) \to 0$ and hence $\gamma(h) \to \mathbf{C}(0)$ implying that $\hat{\gamma}(h)$ must stabilize at some lag $h \geq h_1$. Of course, this requires lot of subjective judgment.

The function chosen involves some parameters that must be estimated from the data. Several methods of estimation have been proposed in the literature (see Cressie (1993), Gaetan and Guyon (2010), and Diggle and Ribeiro (2007)). If one assumes the process is Gaussian, likelihood approach can be used and advocated by Diggle and Ribeiro (2007). The general procedure is to consider the quadratic function

$$Q(\theta) = \sum \left[2\hat{\gamma}(h_i) - 2\gamma(h_i)\right]^2 w_i(\theta),$$

where $w_i(\theta)$ is a weight function chosen a priori. Cressie (1993) suggested one such function, namely,

$$w_i(\theta) = \|N(h)\|/2[2\gamma(h_i, \theta)]^2.$$

The derivation of the above weight function is a bit heuristic, but seems to produce good set of estimates of the concerned parameters, and thus a good Kriging estimator. Alternate weight functions have been suggested recently by Das (2011) in his Ph.D. Thesis submitted to the University of Manchester, UK. The proposed estimators by Das (2011) seem to be more robust against departure from Gaussianity and on the basis of empirical evidence seems to have smaller mean square errors.

1.5. Non-linear quadratic kriging predictor

In the following, we suggest a simple quadratic predictor. Let

$$\widehat{Z}_{quad}(s_0) = \sum_{i=1}^{n_1} a_{1i}(s_{1i})Z(s_{1i}) + \sum_{j=1}^{n_2} b_{2j}(s_{2j})(Z^2(s_{2j}) - \sigma^2) = \beta' \mathbf{Y}(\mathbf{s}),$$

where $\beta' = (a_{11}, a_{12}, \ldots, a_{1n}; b_{21}, b_{22}, \ldots, b_{2n_2})$, $\mathbf{Y}'(\mathbf{s}) = (Z(s_{11}), Z(s_{12}), \ldots, Z(s_{1n_1}); q(s_{21}), \ldots, q(s_{2n_2}))$, and $q(s_{2s_j}) = Z^2(s_{2s_j}) - \sigma^2$, $j = 1, 2, \ldots, n_2$.

The second part of the above coefficient vector and the variable vector $\mathbf{Y}(\mathbf{s})$ corresponds to the nonlinear terms. Here the number of terms in each summation, n_1 and n_2, are chosen on the basis of prior knowledge. We describe briefly how these integers are chosen.

For convenience, we partition each of the two vectors β and $\mathbf{Y}(\mathbf{s})$, into two subvectors, one part corresponding to the linear terms and the other due to nonlinear terms (quadratic terms). Let $\beta^{'} = (\beta_1^{'} \mid \beta_2^{'})$ and $\mathbf{Y}^{'}(\mathbf{s}) = (\mathbf{Z}_{lin}^{'}(\mathbf{s}) \mid \mathbf{Z}_{quad}^{'}(s))$. In order to estimate $\widehat{Z}_{quad}(s_0)$ as usual, we minimize

$$Q_{quad}(\beta) = E(Z(s_0) - \beta^{'}\mathbf{Y}(\mathbf{s}))^2,$$

with respect to β we then obtain the nonlinear Kriging predictor, which can be written as

$$\widehat{Z}_{quad}(s_0) = \sigma_q^{'}(s_0, \mathbf{s})\mathbf{D}^{-1}\mathbf{Y}(\mathbf{s}),$$

where $\sigma_q^{'}(s_0, \mathbf{s}) = \left[EZ(s_0)\mathbf{Z}_{lin}^{'}(\mathbf{s}), EZ(s_0)\mathbf{Z}_{quad}^{'}(\mathbf{s}) \right]$, the matrix is given by the partitioned matrix

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{bmatrix},$$

and the minimum of Q_{quad} is $\sigma^2 - \sigma_q^{'}(s_0, \mathbf{s})\mathbf{D}^{-1}\sigma_q(s_0, \mathbf{s})$. We denote this minimum by Q_{quad}^{min} and the minimum mean square error for the linear predictor by Q_{lin}^{min}. In order to see whether, the inclusion of nonlinear terms gives us a better predictor of $Z(s_0)$, we need to compare these two minima. It is obvious that Q_{quad}^{min} will be less than Q_{lin}^{min}. To construct any statistical test for testing such hypothesis, we need to estimate all the unknown parameters and covariances in these two minima. The sampling distribution

of the statistics, under the null hypothesis, thus obtained need to be investigated. In a classical regression context, the difference (under the null hypothesis that the quadratic terms are absent) will be distributed as a chi-square. In our context, we need to study the distributional properties of these statistics. Some simplifications can be achieved if we set $\mathbf{D}_{12} = \mathbf{D}_{21} = 0$, which can happen if the random process $\{Z(s)\}$ has a symmetric distribution about the mean, even though it is non-Gaussian. If the difference of the minima is significantly large, we would reject the null hypothesis that the linear Kriging is optimal. In order to assess this null hypothesis, we need the sampling distribution of the above difference of the minima, and this needs further investigation. In the classical regression situations, asymptotically, one can show that the difference is distributed as a chi-square under the null hypothesis of linearity with degrees of freedom equal to n_2.

As mentioned before, we need to find the number of terms in each summation, namely, n_1 and n_2. The choice of these numbers depend on the measures of linear and nonlinear dependence among the locations, and these will be discussed below. Here we are presenting new ideas that will be investigated in later publications.

2. Measures for linear dependence and linearity of stationary spatial process

For convenience, and for ease of exposition, we assume $d = 2$ (two-dimensional space), possibly irregular. In spatial literature, several measures were proposed (see, e.g., Cliff and Ord (1981)). In the following, we are guided by the ideas presently used in time series literature. Some of these measures we define now, are similar in spirit to the Moran and Geary's indices (see Gaetan and Guyon (2010) and Schanbenberger and Gotway (2005)). For further discussion on the above measures, see Gaetan and Guyon (2010).

Let us consider the sample $\{Z(s_i); i = 1, 2, \ldots, n\}$, which is a sample from the stationary random field $\{Z(s), s \in \mathbb{R}^2\}$. We discuss our ideas from Kriging point of view. Our object is to estimate $Z(s_0)$, and in doing so, we would like to know, whether the sample is purely random in the sense they are completely independent. If there is a dependence, we would like to know whether it is linear or nonlinear; and if it is nonlinear how far (distance wise) the nonlinearity effect extends. The reason for this reasoning is that as the distances increase the contributions (linear or nonlinear) these terms make on the estimation would decrease. In order to understand these concepts, let us consider a simple example from the area of geo-mining. We would like to estimate the amount of mineral at a location, say, s_0, given the amounts of mineral available at "n" neighboring locations. In this set up, it is obvious that the inclusion of nearest neighbours will give a better estimate and will have more influence than far away locations. In order to locate the nearest neighbours, we proceed as follows: Pictorially, let us represent the locations in the following form (Fig. 1).

Let us divide the entire domain, say Ω, into two disjoint sets $\mathbf{S_1}$ and $\mathbf{S_2}$. Let $\mathbf{S_1}$ contain all the neighbours of s_0,

$$\mathbf{S_1} = \{s_j;\ j = 1, 2, \ldots, n_1;\ \|s_0 - s_j\| \leq \alpha_1\}$$
$$\mathbf{S_2} = \{s_m;\ m, = n_1 + 1, \ldots, M,\ \alpha_1 < \|s_0 - s_m\| \leq \alpha_2\},$$

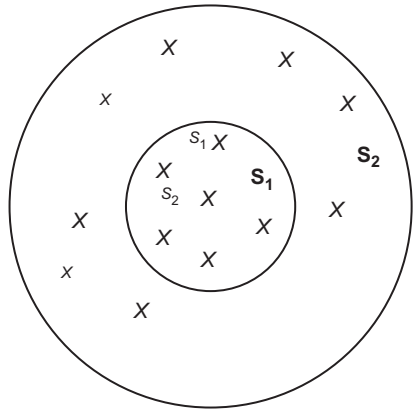

Fig. 1. Random field when $d = 2$.

where $\alpha_1$and α_2 are Euclidean distances chosen a priori. For notational purposes, we denote the elements of the set by $\mathbf{S_1}$ by $\{Z(s_{oj});\ j = 1, 2, \ldots, N_1\}$, and similarly the set $\mathbf{S_2}$ by $\{Z(s_{1j}); j = 1, 2, \ldots, N_2\}$. We note that there are no overlapping elements between these two sets. If we wish to test independence between the elements of these two sets, we consider the correlation coefficient between squares of the elements of these two sets. This will be zero if these two sets are independent. In a similar way, if we wish to test complete randomness within a set, instead of testing correlation between the elements of the two sets, we consider the correlation between the squares of the elements within the set. This methodology is similar to that is followed in the classical nonlinear time series literature, where higher order cumulants are used for testing for Gaussianity and linearity (see Subba Rao and Gabr (1984) and Terdik (1999)). Further, we note that if we consider the set $\mathbf{S_1}$ and if the process is Gaussian, then

$$Cov(Z^2(s_{oj}), Z^2(s_{oj})) = Var(Z^2(s_{oj})) = 2[Var(Z(s_{oj})]^2$$

One can also use the above property for testing Gaussianity. The usual Gaussianity test depends on testing measures of skewness and kurtosis. The sampling distribution of the tests based on the above measures, proposed in the context of spatial processes, need to be investigated. If the elements of the two sets are not nonlinearly dependent, the linear kriging predictor is adequate.

2.1. Intrinsic spatial stationary process

If the process is intrinsically stationary, we have to redefine the above measures as follows. For considering the independence of the elements of the set $\mathbf{S}_1$,we now consider the correlation coefficient between the squares of $(Z(s_{oj} + h) - Z(s_{oj}))$ and $(Z(s_{ok} + h) - Z(s_{ok}))$, for all h. In the light of our earlier discussion, we say the intrinsically stationary process is independent if the correlation coefficient of the squares is zero.

In a similar way, we can define the independence between the two sets of $\mathbf{S}_1$ and $\mathbf{S}_2$ by considering correlation between the squares of the differenced series of the two sets. In a future publication, we wish to consider these measures with possible applications.

3. Models for spatial processes defined on lattices

Most of the finite parameter models proposed for spatial processes, are defined on regular lattices. In the following we briefly consider two-dimensional lattices, though extension to higher dimensions are possible. Let us define the domain $D = \{(i, j); i = 1, 2, \ldots, N; j = 1, 2, \ldots, M\}$. Once again, for convenience, we define $Z(s) = Z(i, j)$, where the sth location is in fact denotes the (i, j)th location defined on a rectangular finite dimensional lattice. There are two widely used models defined on lattices. They are

1. Simultaneous autoregressive models introduced by Whittle (1954) (SAR models)
2. Conditional autoregressive models (CAR) popularized by Besag (1974), Bartlett (1978), and Rozanov (1967).

The conditional autoregressive models are also known as random field models (Rozanov, 1967). Besag (1974) considered the estimation of the CAR models using coding technique and likelihood approach based on coding method. A frequency domain approach for the estimation of Markov random field models have been discussed by Yuan (1989), and Yuan and Subba Rao (1993), and the frequency domain method used for these models can also be generalized for all dimensional lattices and does not depend on the Gaussianity assumption.

3.1. Simultaneous autoregressive models (SAR)

The SAR models defined by Whittle (1954) are similar to the classical time series AR models. Here the random process $Z(s)$, $s \in \mathbb{Z}^2$ is assumed to satisfy a finite difference equation of the form

$$Z(s) = \sum_{u \in S} a(u) Z(s + u) + e(s)$$

where the set S is set of all nearest neighbors of the location $u \in S$ such that u $\neq$ 0. The errors $\{e(s)\}$ are often assumed to be a sequence of independent, identically distributed random variables, possibly each having a normal distribution with mean zero and variance σ_e^2. The estimation of SAR models have been considered by Ord (1975).

3.2. Conditional autoregressive models (CAR)

In this formulation, the model is stated in terms of conditional expectations, unlike SAR models. The model is defined as follows:

$$E[Z(s)|Z(s + u), u \in S, u \neq 0] = \sum_{u \in S} a(u) Z(s + u)$$

which can also be written as $Z(s) = \Sigma_{u \in S} a(u) Z(s + u) + \eta(s)$. The models given above may look alike, but the fundamental difference is in the SAR model: the sequence of errors are not independent of $\Sigma_{u \in S} a(u) Z(s + u)$, but in the case of CAR, because of

our definition through the conditional expectations, the errors $\{\eta(s)\}$ are independent under the Gaussianity set up as described by Besag (1974, 1975) and Ord (1975). As pointed out by Besag (1974) and Ord (1975), the ordinary least squares estimates in the case of SAR models are not consistent, where as in the case of CAR, they are consistent. This is because of the fact that in the case of CAR models, the conditional expectation $E(\eta(s) \mid \Sigma_{u\in S} a(u)Z(s+u)) = 0$. Below we consider the frequency domain approach for the estimation of CAR models (Markov Random field models). For details, we refer to Yuan (1989) and Yuan and Subba Rao (1993).

4. Frequency domain approach for the estimation of CAR models

In order to present the details, we need to describe briefly the spectral representations of stationary random processes and the definition of the spectral density functions. Since we are considering processes defined on a two-dimensional lattice, we denote the random process as $Z(t)$, where $t \in \mathbb{Z}^2$. Here $t = (t_1, t_2)$. Let $Z(t)$ be a zero mean, second-order stationary process admitting the spectral representation

$$\mathbf{Z}(t) = \int\int_{-\pi}^{\pi} \exp(it.\omega) dZ_x(\omega)$$

$t.\omega = t_1\omega_1 + t_2\omega_2$ and $dZ_X(\omega) = dZ(\omega_1, \omega_2)$. Let us define the covariance function of lag difference "h", $C(h)$ by

$$C(h) = \int\int_{-\pi}^{\pi} \exp(ih.\omega) f(\omega) d\omega,$$

and the fundamental frequency range is $|\omega_i| \le \pi$, $i = 1, 2$. Here the $f(\omega)$ is defined as the second-order spectral density function of the random process $Z(t)$. By inverting the above, we get

$$f(\omega) = \frac{1}{(2\pi)^2} \sum_h \exp(-ih.\omega) C(h),$$

If we define a real-valued Hilbert space spanned by the random variable $\{Z(t)\}$, with norm $Var(Z(t)) = E(Z(t))^2$, inner product $Cov(Z(t_1), Z(t_2))$, a corresponding sub-Hilbert space spanned by $\{Z(t+u);\ u\epsilon S,\ u \neq 0\}$, then we obtain, for all $v\epsilon S$, $v \neq 0$,

$$E[Z(s) - \sum_{u\epsilon S} a(u)Z(s+u)]Z(v) = 0,$$

as $\Sigma a(u)Z(s+u)$ is nothing but the projection onto the sub-Hilbert space of $\{Z(s)\}$. The above equation implies, in view of stationarity

$$C(v-s) - \sum_{u\epsilon S} a(u)C(v-s-u) = 0,$$

In the above equation, if we substitute the spectral representations, we obtain

$$\int\int_{-\pi}^{\pi} \exp(i(v-s).\omega) f(\omega) \left[1 - \sum_{u \epsilon S} a(u) \exp(-iu.\omega)\right] d\omega = 0$$

which implies,

$$f(\omega) = k \left[1 - \sum_{u \epsilon S} a(u) \exp(-iu.\omega)\right]^{-1}.$$

If $f(\omega) > 0$, $f^{-1}(\omega) = K\left[1 - \sum_{u \epsilon S} a(u) \exp(-iu.\omega)\right]$. This shows, that in the case of CAR models, defined on lattices, the inverse spectrum is linear in terms of the coefficients $\{a(u)\}$. This suggests that, if we have a Fourier expansion for the inverse spectrum, the Fourier coefficients of the expansion, are in fact proportional to the coefficients $\{a(u)\}$. In his Ph.D. Thesis, Yuan (1989), considered the estimation of the parameters of the model using the above observation and also studied their sampling properties. If the coefficients are symmetric in the neighbourhood of $u = 0$, then

$$f^{-1}(\omega) = K \left[1 - \sum_{u \epsilon S} a(u) cos\,(u.\omega)\right].$$

The order of the neighbourhood can also be estimated, by studying the properties of the inverse autocovariances as they tend to zero as the lags go beyond the defined neighbourhood.

It is important to note that the above estimation, does not depend on the assumption of Gaussianity, unlike the estimation procedures considered earlier.

5. Spatio-temporal processes

Let $Z(s,t); s \in \mathbb{R}^d, t \epsilon Z$ be a zero mean, spatially and temporally stationary process, by this we mean, $E(Z(s,t)) = 0$, $Var(Z(s,t)) = \sigma^2$, $Cov(Z(s,t), Z(s+h, t+u)) = C(h,u)$, $h \in \mathbb{R}^d$, $u \in Z$. Here, "h" denotes the spatial lag and "u" denotes the temporal lag. If $C(h,u) = R(\|h\|, u)$, we say the process is spatially isotropic. The literature on spatio-temporal processes is not as rich as in spatial processes, but a recent book by Cressie and Wikle (2011) may fill this gap. A special class of processes, widely known as "separable processes" have been proposed (see Cressie and Huang (1999), Fuentes (2006), Gneiting (2002), and Ma (2004)) recently. As far as the authors are aware, literature on finding suitable finite parametric models, methods of estimation, etc., in the case of spatio-temporal processes are still at an early stage. In this section, we will try to define some suitable new models, and their estimation, etc., will be considered in future publications.

We say the random process $Z(s,t)$ is separable if the covariance function can be represented as product of two covariances as $C(h,u) = C_1(h)C_2(u)$. This implies that

the second-order spectral density function defined for spatial processes $Z(s,t)$ can be written as a product of two second-order spectra: one due to the spatial processes and the other one is due to the temporal processes.

Consider a separable process with covariance function $C(h,u) = C_1(h)C_2(u)$. Taking two dimensional Fourier transforms both sides(over h and u) we get,

$$f(\lambda,\omega) = f_1(\lambda) f_2(\omega),$$

where $|\omega| \leq \pi$, $|\lambda| \leq \pi^d$, $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_d)$. From the above equation, we get after taking logarithms.

$$\ln f(\lambda,\omega) = \ln f_1(\lambda) + \ln f_2(\omega),$$

which is additive. This decomposition provides us a test for separability of the processes (see Fuentes (2006)), in fact this idea of separation of spectral density functions was used by Priestley and Subba Rao (1969) for constructing a statistical test for second order stationarity.

5.1. Models for spatio-temporal processes

Here, we briefly describe some ideas that we believe to be new for defining spatio-temporal processes. The models we suggest below are similar to SAR and CAR models suggested earlier for purely spatial processes. We define in terms of frequency domain.

5.2. Conditional autoregressive spatio-temporal (CAST) models

Let us assume that $Z(s_j,t)$, $s_j \in \mathbb{Z}^2, : t \in \mathbb{Z}$ is the random process defined on a finite lattice (for each t) $i = 1,2,\ldots,M_1$; $j = 1,2,\ldots,M_2$, and $t = 1,2,\ldots,N$. We are considering a finite dimensional lattice with $M_1 M_2$ locations. Let us define the discrete finite Fourier transform of the time series observed at a specific location, say s_i, by

$$J_{s_j}(\omega_k) = \frac{1}{\sqrt{2\pi N}} \sum_t Z(s_j,t) \exp(-it\omega_k),$$

where $\omega_k = 2\pi k/N$, $k = 0, \pm 1, \pm 2, \ldots, \pm [N/2]$. It is well known that if the process is second-order temporally stationary, the discrete Fourier transforms $J_{s_j}(\omega_k)$ and $J_{s_j}(\omega_{k'})$ are asymptotically uncorrelated over distinct frequencies if the process is stationary and asymptotically independent if the process is a Gaussian stationary process. Further, under the assumption that the mean is zero, we have

$$E(J_{s_j}(\omega)) = E\left[\frac{1}{\sqrt{2\pi \mathbf{N}}} \sum_t Z(s_j,t) exp(-it\omega)\right] = 0,$$

$$Var(J_{s_j}(\omega_k)) \simeq f_{s_j}(\omega_k).$$

If the process is Gaussian, the discrete Fourier transforms are complex Gaussian. Now following Rozanov (1967), we define a Hilbert space defined by complex random variables (the discrete Fourier transforms of the processes at all the locations at a given

frequency ω), and a suitable sub-Hilbert space spanned by a subset. For convenience of exposition, we denote each location by s_i, where $i = 1, 2, \ldots, M_1 M_2$, i.e., all the coordinates corresponding to the entire lattice. The inner product of the Hilbert space spanned by all the discrete Fourier transforms is the covariance between the discrete Fourier transforms $Cov(J_{s_i}(\omega), J_{s_j}(\omega)) = f_{s_i s_j}(\omega)$, where $f_{s_i s_j}(\omega)$ is the cross-spectral density function of the spatio-temporal processes defined at two locations s_i and s_j and in view of our assumption of spatial stationarity is a function of the lag difference $s_i - s_j$ only. If the process is assumed to be isotropic, the lag difference is the Euclidean distance "h" only. We note when $i = j$, $h = 0$, then the cross-spectral density function is nothing but the spectral density function at a given location (i or j), and in view of stationarity, it is same for all locations. For a fixed frequency ω, let us define the conditional expectation

$$E[J_s(\omega) \mid J_{s+u}(\omega); u \epsilon S, u \neq 0] = \sum_{u \in S} \beta(u) J_{s+u}(\omega),$$

where the coefficients $\{\beta(u)\}$ can be complex valued. From the above conditional expectation, we get, for all $v \neq 0$,

$$E\left[J_s(\omega) - \sum_{u \in S} \beta(u) J_{s+u}(\omega) \right] J_v^*(\omega) = 0,$$

which gives a set of Yule–Walker-type equations in terms of the spectral density functions, which can be solved to obtain the coefficients $\{\beta(u)\}$. The above equations reduce to (for a given ω),

$$f_{v-s}(\omega) - \sum_{u \in S} \beta(u) f_{v-s-u}(\omega) = 0,$$

for all $s = 1, 2, \ldots, M_1 M_2$. In the above summations, we should note that u takes all values belonging to the neighbourhood of S except $u = 0$.

The estimation of the above coefficients and the sampling properties will be discussed in future publications. We can formally rewrite the above conditional expectation in the form of a model using the discrete Fourier transforms;

$$J_s(\omega) = \sum_u \beta(u) J_{s+u}(\omega) + J_{s,e}(\omega),$$

where $\{J_{s,e}(\omega)\}$ is a sequence of complex-valued Gaussian random variables (for a fixed ω) independent of each other overall locations $\{s\}$. In view of our definition through the conditional expectations, the two terms on the right-hand side of the above equation are independent. This is similar to CAR models considered by Besag (1974) and others for purely spatial processes. The estimation of the parameters and statistical inference associated with the estimates will be considered elsewhere.

5.3. Simultaneous autoregressive spatio-temporal models (SAST)

We give here a frequency domain formulation of the SAST models. Ali (1979) generalized SAR models to time series data defined on lattices. Following the notation of

Ali (1979), let us define $Z(l,m;t)$ as the stationary time series defined at the location (l,m) for each "t". The SAR model is defined as

$$Z(l,m;t) = \sum_i \sum_j \sum_{k=1}^{p} \phi\,(i,j;k)\, Z(l-i, m-j; t-k) + e(l,m;t),$$

The estimation of the parameters was considered by Ali (1979). The choice of the orders, nearest neighbourhood S over which the model is defined have to be considered, and as far as the authors are aware, no studies are available dealing with these problems. Now we define a frequency domain version of SAST models.

5.4. *Frequency domain SAST*

As defined earlier when discussing CAST models, we define the discrete Fourier transforms at each location as before. Now define the SAST model

$$J_s(\omega) = \sum_{u \epsilon S} \gamma(u) J_{s+u}(\omega) + J_{s,\eta}(\omega)$$

where $\{J_{\mathbf{s},\eta}(\omega)\}$ is a sequence of independent complex random variables, independent over s for a fixed ω. The estimation, sampling properties of the estimators and applications of these need to be investigated. Some of these will be considered elsewhere. As stated earlier when discussing CAR and SAR models for spatial processes, the two terms on the right-hand side of the above model are not orthogonal.

6. Multivariate AR and STAR models

In the following discussions, we do not need to assume that the space is a lattice (either two dimensions or more than two). We assume here that $s \in \mathbb{R}^d$. Let $Z(s_i,t)$; $i = 1,2,\ldots,n$, $t = 1,2,\ldots,N$. In other words, we have a time series of length N at each of the n locations. Define the vector, for each t,

$$\mathbf{Z}'(t) = (Z(s_1,t), Z(s_2,t), \ldots, Z(s_n,t)).$$

Then one can consider the above vector process as a multivariate stationary time series and use the classical multivariate ARMA models for modeling the spatio-temporal data. The multivariate $ARMA(p,q)$ can be written as

$$\mathbf{Z}(t) = \sum_{j=1}^{p} A(j)\mathbf{Z}(t-j) + \sum_{k=0}^{q} B(k)\mathbf{e}(t-k),$$

where $B(0) = I$, and the random vectors $\{\mathbf{e}(t)\}$ are assumed to be mutually independent, identically distributed random vectors (usually assumed to be Gaussian, although not always necessary) with mean zero and variance covariance matrix $\mathbf{G}$. The matrix

coefficients need to satisfy some conditions in order to satisfy the conditions of second-order stationarity and invertibility (see Priestley (1981) and Lutkepohl (2008)). These coefficients can be estimated using maximum likelihood estimation methodology. In the above, we assumed that the mean of the spatio-temporal process $\{\mathbf{Z}(t)\}$ zero. Suppose the random process, say $\{\mathbf{Y}(t)\}$**,** represents the original process with non-zero mean, say $\mathbf{X}(t, \vartheta)$. Then we can write $\{\mathbf{Y}(t)\}$ as

$$\mathbf{Y}(t) = \mathbf{X}(\mathbf{t}, \vartheta) + \mathbf{Z}(t),\ t = 1, 2, 3, \ldots, N$$

where $\{\mathbf{Z}(t)\}$ is a zero mean spatio-temporal processes. Usually, the mean component $\mathbf{X}(t, \vartheta)$ is a function (linear or nonlinear) of some co-variates.

In many real applications, such as climatology and physical sciences, the mean component can be a function of time (polynomial in time) representing the trend, seasonality, or both. Such models have been recently considered by Terdik et al. (2007), Subba Rao and Terdik (2006), and Gao and Subba Rao (2010). Terdik et al. (2007) and Subba Rao and Terdik (2006) have developed frequency domain methods for estimating the nonlinear regression parameters and studied their sampling properties. As stated earlier, one of the important objects in environmental time series, such as global temperatures, etc., is to detect whether there is a global warming and if so, by how much. Hughes et al. (2007) considered the minimum and maximum monthly temperatures at Faraday Station, Antarctic Peninsula and estimated the amount of increase by estimating the trend under the assumption that the errors are correlated and admit an ARMA model with innovations having an extreme value distribution. In situations where such assumptions are not possible, we believe the frequency domain methods developed by Terdik et al. (2007) and Subba Rao and Terdik (2006) are very useful. One disadvantage of using multivariate ARMA models given above in spatial context is that as the number of locations n increases, the dimensions of the coefficients increase. With this increase of dimensionality, we may encounter severe problems in the maximization of the likelihood functions. As an alternative to these, a class of models known as space–time autoregressive models (STARMA) were proposed by Pfeifer and Deutsch (1980). These models are parsimonious that not only include lag terms, but also take into account the "distances" between the locations. Besides the above papers, we also refer to recent papers by Subba Rao and Antunes (2004), and Antunes and Subba Rao (2006). The dependence between the "n" locations is characterized through a sequence of $n \times n$ matrices, specified by the practitioners before hand. Suppose we denote the l^{th} order spatial dependence matrix by $\mathbf{W}^l$, $(l = 1, 2, \ldots, n)$, then the STARMA$(p_{\lambda_1\lambda_2\ldots\lambda_p}, : q_{m_1m_2\ldots m_q})$ is defined as

$$\mathbf{Z}(t) = -\sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\varphi_{kl}\mathbf{W}^l Z(t-k) + e(t) + \sum_{k=1}^{q}\sum_{l=0}^{m_k}\theta_{kl}\mathbf{W}^l e(t-k)$$

where the coefficients $\{\varphi_{kl};\theta_{kl}\}$ are scalars. Here the random errors $\{e(t)\}$ are assumed to be independent normal vectors with zero mean and variance covariance matrix $\mathbf{G}$. Since the weighting matrices are chosen before hand, the problem is to estimate the parameters of the above equation given that we have T consecutive observations at each of the "n" locations. The identification of STARMA are done using space–time

autocorrelation coefficients and partial autocorrelation coefficients (see Pfeifer and Deutsch (1980)) similar to the classical time series. The space–time autocovariance function of lag s between lth and kth-order neighbors is defined by

$$\gamma_{lk}(s) = E\left\{\frac{(\mathbf{W}^l\mathbf{Z}(t))'(\mathbf{W}^k\mathbf{Z}(t+s))}{N}\right\}$$

and the autocorrelation coefficient can be defined as

$$\rho_{lk}(s) = \frac{\gamma_{lk}(s)}{\{\gamma_{ll}(0)\gamma_{kk}(0)\}^{\frac{1}{2}}}.$$

These coefficients can be estimated since the weighting matrices are known. As in time series, the partial space–time autocorrelation coefficients become zero beyond the order of the STAR model, and space–time autocorrelation coefficients become zero for STMA models (see Subba Rao and Antunes (2004)).

Under the assumption of Gaussianity and $\mathbf{G} = \sigma^2\mathbf{I}$,we can show that the maximization of the log-likelihood function is same as the minimization of

$$S(\varphi, \theta) = \sum_{i=1}^{N}\sum_{t=1}^{T} e_i^2(t)$$

with respect to the parameter vector (φ, θ). In order to implement the minimization algorithm, we need the residuals $\{e_i(t)\}$. Subba Rao and Antunes (2004) have suggested using a procedure similar to Hannan and Rissanen (1982). The estimates obtained did converge. Subba Rao and Antunes (2004) fitted the above models to monthly mean temperatures, recorded in Celsius scale, at nine meteorological stations around the United Kingdom. The data was accessed through the Web site of the LDEO/IRI Data library found in http://rainbow.ldeo.columbia.edu/. There are 223 observations available for the nine sites from January 1951 through to August 1969. In order to save space, we will not reproduce the results as they can be found in Subba Rao and Antunes (2004). The object in this chapter was not only to illustrate the method of estimation of STARMA models, but also to compare the forecasting performance of the STARMA models with univariate ARMA models fitted using standard software packages and methodology. At least in this case, it was found that STARMA models gave better forecast than univariate models, and the obvious reason is that in the space–time modeling we have not only taken temporal correlation into account, but also spatial correlation. Another advantage is that these STARMA models have less parameters to estimate.

Since multivariate ARMA models STARMA models look similar, a natural question one would consider is whether one model is nested in the other model. The question of testing of non-nested hypotheses was considered by Cox (1961) in the case of independent, identically distributed random variables. Except for Walker (1967), not much is known regarding testing non-nested hypotheses in the case of time series. Now consider a special case of STAR model where only a first-order dependence is taken

into account. The model considered by Antunes and Subba Rao (2006) is of the form

$$\mathbf{Z}(t) + \sum_{k=1}^{p} (\phi_k \mathbf{I}_n + \psi_k \mathbf{W})\, \mathbf{Z}(t-k) = \varepsilon(t),$$

where the coefficients $\{\phi_{\mathbf{j}}, \psi_{\mathbf{j}}; j = 1, 2, \ldots, p\}$ are scalars and the first-order weighting matrix $\mathbf{W}$ is assumed to be known. Now consider a n-variate multivariate AR model of order q of the form

$$\mathbf{Z}(t) + \sum_{j=1}^{q} \mathbf{A}(j)\mathbf{Z}(t-j) = \nu(t),$$

where the random vectors $\{\varepsilon(t)\}$ and $\{\nu(t)\}$ are assumed i.i.d, each having a zero mean multivariate normal distribution with variance covariance matrices Σ_ε and Σ_ν, respectively. Antunes and Subba Rao (2006) have shown that STAR model is nested within multivariate AR model only if the orders p and q are equal. The above two models were fitted to 500 hourly carbon monoxide atmospheric concentrations in four different locations in London beginning January 01, 2004. The locations are Bloomsbury, Hillington, Marylebone Road, and Westminster). In both cases, the orders were chosen using Quinn (1980) information criterion. The data considered was logarithmically transformed and deseasonalized. Using the above criterion, it was found that an order two STAR model found to be appropriate, and for the multivariate case it was found AR model of order 3 was found to be most suitable. Though the residual variance–covariance matrices in both cases look similar, one can easily see that the STAR model has less number of parameters compared to Multivariate AR models. If one decides to choose a parsimonious model, in this case at least, one should choose a STAR model. For further discussion, see Antunes and Subba Rao (2006).

6.1. Non linear space–time models (space–time bilinear models)

In time series context, Granger and Andersen (1978), Subba Rao (1977, 1981), Subba Rao and Gabr (1984), and Terdik (1999) have defined a class of nonlinear time series models, and now they are widely known as bilinear models. Let us briefly consider the univariate case.

Let $\{X_t\}$ be a discrete parameter time series. We say the time series is a bilinear process, if the time series satisfies the difference equation

$$X_t + \sum_{j=1}^{p} a_j X_{t-j} = \sum_{j=0}^{r} c_j e(t-j) + \sum_{l=1}^{m}\sum_{k=1}^{q} b_{lk} X_{t-l} e(t-k),$$

where $c_0 = 1$ and $\{e_t\}$ are independent, identically distributed random variables with mean zero and variance σ_e^2. We define the above models as $BL(p, r, m, q)$. The properties of $BL(p, 0, p, 1)$ have been investigated by Subba Rao (1981), and Terdik (1999).

The estimation of parameters of bilinear models and subset bilinear models have also been considered in the above publications. These univariate bilinear models have been extended to multivariate situations by Stensholt and Tjøstheim (1987), and Subba Rao and Wong (1999). An interesting property of bilinear models is that the general solution of these equations (under certain conditions) have Volterra representation, and one can obtain explicit expressions for higher order cumulants (see Terdik (1999)). As in linear AR models, bilinear models also satisfy Yule–Walker type difference equations in terms of higher order cumulants. Another interesting feature of the bilinear models are that the conditional mean of the process defined by these models is dependent (nonlinearly) on the past set of data and conditional variance is constant. This is in contrast to ARCH and GARCH models where the conditional mean is constant (usually assumed to be zero), but the conditional variance is changing and dependent on the past data. In view of this property the ARCH and GARCH models are used for representing the stochastic volatility in the financial markets.

Dai and Billard (1998) have extended the bilinear models described above to spatial situations. Following earlier notation, we let $\mathbf{Z}(t)$ to represent the time series at time "t" at all n locations. The space–time bilinear model (Dai and Billard, 1998) is

$$\mathbf{Z}(t) = \sum_{i=1}^{p}\sum_{m=0}^{\lambda_i} \phi_m^i \mathbf{W}^m \mathbf{Z}(t-i) + \sum_{j=1}^{q}\sum_{n=0}^{\eta_j} \theta_n^j \mathbf{W}^n \mathbf{e}(t-j) + \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{m=0}^{\xi_i}\sum_{n=0}^{\mu_j} \beta_{mn}^{ij} [\mathbf{W}^m \mathbf{Z}(t-i)] \# [\mathbf{W}^n \mathbf{e}(t-j)] + \mathbf{e}(t),$$

where $\{\mathbf{e}(t)\}$ assumed to be i.i.d random vectors and $\mathbf{A}\#\mathbf{B} = \mathbf{C} = (c_{ij})$, $c_{\mathbf{ij}} = a_{ij} b_{ij}$.

The conditions of stationarity, etc., were considered by Dai and Billard (1998). As far as we are aware, these models have not been used in the analysis of real data. It is important to study the importance of these models in physical and biological sciences. In the context of climatology, it would be useful to see how these models can be used to estimate an observation at a known location given the data at other locations. We hope these problems will be considered in future publications.

Concluding Remarks

In this review, our object is to show how some of the well-known techniques such as linear prediction (Kriging) methodology for the estimation of an unknown observation at a known location can be extended to nonlinear situations and also known spatial process models to include temporal dimension. The new spatio-temporal models defined here are based on discrete Fourier transforms. These methods here defined led us to several open and interesting problems, and we hope to consider some of these. Also we hope the readers will also consider these.

The applicability of these methods and new techniques need to be applied and tested in the case of real data.

Acknowledgements

We thank the reviewers for their constructive and useful comments which improved the presentation.

This work is partially supported by TÁMOP 4.2.1./B-09/1/KONV-2010-0007/IK/IT project. The project is implemented through the New Hungary Development Plan co-financed by the European Social Fund, and the European Regional Development Fund.

References

Ali, M.M., 1979. Analysis of stationary spatial-temporal processes: estimation and Prediction. Biometrika 66, 513–518.

Antunes, A.M., Subba Rao, T., 2006. On hypotheses testing for the selection of spatio-temporal models. J. Time Ser. Anal. 27, 767–791.

Bartlett, M.S., 1978. Nearest neighbour models in the analysis of field experiments. J. R. Stat. Soc. Ser. B 40, 147–174.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B 36, 192–225.

Besag, J., 1975. Spatial analysis of non-lattice data. Statistician 24, 179–195.

Cliff, A., Ord, J.K., 1981. Spatial Processes: Models and Applications. Pion, London.

Cox, D., 1961. Tests of separate families of hypotheses. Proceedings of the 4th Berkeley Symposium, Berkeley, pp. 105–123.

Cressie, N., 1993. Statistics for Spatial Data. John Wiley, New York.

Cressie, N., Huang, H.-C., 1999. Classes of non-separable, spatio-temporal stationary covariance functions. J. Am. Stat. Assoc. 94, 330–340.

Cressie, N., Wikle, C.K., 2011. Statistics for Spatio-Temporal Data. Wiley Series in Probability and Statistics, Hoboken, NJ.

Dai, Y., Billard, L., 1998. A space time bilinear model and its identification. J. Time Ser. Anal. 19, 657–679.

Das, S., 2011. Statistical estimation of variogram and covariance parameters of spatial and spatio-temporal random processes. Unpublished Ph.D. Thesis submitted to the University of Manchester, United Kingdom.

Diggle, P.J., Ribeiro, P.J., 2007. Model-Based Geostatistics. Springer-Verlag, New York, NY.

Fuentes, M., 2006. Testing for separability of spatio-temporal covariance functions. J. Stat. Plan. Inference 13, 447–466.

Gaetan, C., Guyon, X., 2010. Spatial Statistics and Modeling. Springer-Verlag, New York, NY.

Gao, X., Subba Rao, T., 2010. Regression models with STARMA models: an application to the study of temperature variations in the Antarctic Peninsula. In: Martin, T.W., Ashish, S. (Eds.), Festchrift for S. R. Jammalamadaka. Springer-Verlag, pp. 27–50.

Gneiting, T., 2002. Non-separable stationary covariance functions for space-time. J. Am. Stat.Assoc. 97, 590–600.

Granger, C.W.J., Andersen, A.P., 1978. An introduction to Bilinear time series models. Vandenhoek and Ruprecht, Gottingen.

Hannan, E.J., Rissanen, J., 1982. Recursive estimation of mixed autoregressive moving average models. Biometrika 69, 81–94.

Hughes, G., Subba Rao, S., Subba Rao, T., 2007. Statistical analysis and time series models for minimum/maximum temperatures in the Antarctic Peninsula. Proc. R. Soc. Ser A 461, 241–259.

Lahiri, S.N., Lee, Y., Cressie, N., 2002. On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. J. Stat. Plan. Inference. 103, 65–85.

Lutkepohl, H., 2008. New Introduction to Multiple Time Series Analysis. Springer-Verlag, New York, NY.

Ma, C., 2004. Spatial autoregression and related spatio-temporal models. J. Mult. Anal. 88, 152–162.

Matern, B., 1986. Spatial variation. Lecture Notes in Statistics. Springer-Verlag, New York, NY.

Ord, J.K., 1975. Estimation methods for models of spatial interaction. J. Am. Stat. Assoc. 70, 120–126.

Pfeifer, P., Deutsch, S., 1980. A three stage interacting procedure for space-time modeling. Technometrics 22, 35–47.

Priestley, M.B., 1981. Spectral Analysis of Time Series. Academic Press, New York, NY.
Priestley, M.B., Subba Rao, T., 1969. A test for non-stationarity of time series. J. Roy. Stat. Soc. Ser B 31, 140–149.
Quinn, B., 1980. Order determination for a multivariate autoregression. J. Roy. Stat. Soc. B. 42, 182–185.
Rozanov, Y.A., 1967. On the Gaussian homogeneous fields with given conditional distributions. Theory Prob. Appl. 12, 381–391.
Schanbenberger, O., Gotway, C.A., 2005. Statistical Methods for Spatial Data Analysis. Texts in Statistical Series. Chapman Hall/CRC, Boca Raton.
Stensholt, B.K., Tjøstheim, D., 1987. Multiple bilinear time series models. J. Time Ser. Anal. 8, 221–233.
Subba Rao, T., 1977. On the estimation of bilinear time series models. Bull Int. Stat. Inst, vol 41, 139–140 (Paper presented at the 41st session of ISI, New Delhi).
Subba Rao, T., 1981. On the theory of bilinear time series models. J. Roy. Stat. Soc. B 43, 244–255.
Subba Rao.T and Gabr M.M. (1984) An introduction to biopectral analysis and linear time series moduls vol 24. Springer-verlag. New York.
Subba Rao, T., Antunes, A., 2004. Spatio-temporal modelling of temperature time series-a comparative study. In: Schonberg, F., Brillinger, D.R., Robinson, E. (Eds.), Time Series analysis and Applications to Geophysical Systems, vol. 139. IMA Publications. Springer-Verlag, pp. 105–122.
Subba Rao, T., Terdik, Gy., 2006. Multivariate non-linear regression with applications. Lecture Notes in Statistics, No. 187. In: Bertail, P., Doukhan, P., Soulier, P. (Eds.), Dependence in Probability and Statistics, Springer Verlag, New York, pp. 431–470.
Subba Rao, T., Wong, W., 1999. Some contributions to multivariate bilinear time series models. In: Ghosh, S. (Ed.), Asymptotics, Nonparametrics and Time Series. Marcell Dekker, New York.
Terdik, Gy., 1999. Bilinear Stochastic Models and Related Problems of Nonlinear Time Series Analysis: A Frequency Domain Approach. Lecture Notes in Statistics, vol. 142. Springer, New York, NY.
Terdik, Gy., Subba Rao, T., Jammalamadaka Rao, S., 2007. On multivariate nonlinear regression models with stationary correlated errors. J. Stat. Plan. Inference, vol 137, 3793–3814.
Yuan, J., 1989. Spectral analysis of multidimensional stationary process with applications with applications in image processing. Unpublished Ph.D. Thesis submitted to the University of Manchester Institute of Science and Technology (UMIST), U.K. Now It is the University of Manchester.
Yuan, J., Subba Rao, T., 1993. Spectral estimation for random fields with applications to Markov modeling and texture classification. In: Chellappa, R., Jain, A. (Eds.), Markov Random fields-Theory and Applications. Academic Press, pp. 179–209.
Walker, A.M., 1967. Some tests of separate families of hypotheses in time series analysis. Biometrika 54, 39–68.
Whittle, P., 1954. On stationary processes in the plane. Biometrika 49, 305–314.

Part VIII: Continuous Time Series

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00019-3

Lévy-Driven Time Series Models for Financial Data

Peter Brockwell[1] *and Alexander Lindner*[2]

[1]*Departments of Statistics, Colorado State University and Columbia University*
[2]*Institut für Mathematische Stochastik, TU Braunschweig*

Abstract

The ARCH and GARCH models of Engle (1982) and Bollerslev (1986) respectively have had great success in the modeling of financial time series. Discrete-time stochastic volatility models have also been found to be very useful in representing the time variation of volatility observed in such data. In this review, we discuss Lévy-driven continuous-time versions of these processes and some related inference questions.

Keywords: Lévy process, Lévy-driven CARMA process, stochastic volatility, COGARCH process, generalized Ornstein–Uhlenbeck process.

AMS Classification: 62M10, 60H10, 91G70

1. Introduction

The study of time series with continuous-time parameter received great impetus from the very successful application of such processes in theoretical finance, particularly with the work of Black, Scholes, and Merton on the pricing of options and the subsequent explosive growth of financial mathematics. An excellent recent overview of financial time series can be found in the book edited by Andersen et al. (2009). For further applications of Lévy processes in finance, we recommend also the books of Cont and Tankov (2004) and Schoutens (2003).

In this article, we focus attention on some financial time series driven by Lévy processes, the essential properties of which are introduced in Section 2. Lévy processes play a central role for several reasons, one being that their sample paths are not restricted to be continuous and another that the distributions of their increments can be any of the very large class of infinitely divisible distributions. In Section 3,

we discuss Lévy-driven continuous-time autoregressive moving average (CARMA) processes that play a role in continuous time analogous to their discrete-time counterparts and introduce the stationary Lévy-driven Ornstein–Uhlenbeck process as a special case. Section 4 deals with the celebrated continuous-time stochastic volatility model of Barndorff-Nielsen and Shephard (2001) in which the volatility is a stationary Ornstein–Uhlenbeck process driven by a *subordinator* (a Lévy process with nondecreasing sample paths). In Section 5, we consider an extended version of the Barndorff-Nielsen–Shephard model in which the volatility is a non-negative CARMA process and discuss parameter estimation for the volatility based on realized integrated volatility. In Section 6, we introduce the generalized Ornstein–Uhlenbeck (GOU) process, which generalizes the Ornstein–Uhlenbeck process in a direction different from that of the CARMA process. Finally in Section 7, we discuss the COGARCH(1,1) process of Klüppelberg et al. (2004), in which the volatility is a GOU process, and the higher order COGARCH(p,q) process of Brockwell et al. (2006).

2. Lévy processes

A *Lévy process* with values in $\mathbb{R}^d$ $(d \in \mathbb{N})$ defined on a probability space $(\Omega, \mathcal{F}, P)$ is a stochastic process $M = (M_t)_{t\geq 0}$, $M_t : \Omega \to \mathbb{R}^d$ with independent and stationary increments such that $M_0 = 0$ almost surely and the sample paths are almost surely right continuous with finite left limits. By *independent increments*, we mean that for every $n \in \mathbb{N}$ and $0 \leq t_0 < t_1 < \cdots < t_n$, the random variables M_{t_0}, $M_{t_1} - M_{t_0}$, and $M_{t_2} - M_{t_1}, \ldots, M_{t_n} - M_{t_{n-1}}$ are independent, and by *stationary increments,* we mean that $M_{s+t} - M_s$ has the same distribution as M_t for every $s, t \geq 0$. We refer to the books by Applebaum (2004), Bertoin (1996), Kyprianou (2006), and Sato (1999) for further information about Lévy processes, in which the proofs for the results stated in this section can also be found.

Elementary examples of Lévy processes $M = (M_t)_{t\geq 0}$ with values in $\mathbb{R}^d$ include linear deterministic processes of the form $M_t = bt$, where $b \in \mathbb{R}^d$, d-dimensional Brownian motion and d-dimensional compound Poisson processes. If $M = (M_t)_{t\geq 0}$ is any Lévy process, then for all t the distribution of M_t is characterized by a unique triplet (A_M, ν_M, γ_M) consisting of a symmetric non-negative $d \times d$ matrix A_M, a measure ν_M on $\mathbb{R}^d$ satisfying $\nu_M(\{0\}) = 0$ and $\int_{\mathbb{R}^d} \min\{|x|^2, 1\}\, \nu_M(\mathrm{d}x) < \infty$ and a constant $\gamma_M \in \mathbb{R}^d$. This triplet determines the characteristic function of M_t via the Lévy–Khintchine formula,

$$E\mathrm{e}^{i\langle M_t, z\rangle} = \exp\left\{ i\langle \gamma_M, z\rangle - \frac{1}{2}\langle z, A_M z\rangle + \int_{\mathbb{R}^d} (\mathrm{e}^{i\langle z,x\rangle} - 1 - i\langle z, x\rangle \mathbf{1}_{\{|x|\leq 1\}})\, \nu_M(\mathrm{d}x) \right\} \tag{1}$$

for $z \in \mathbb{R}^d$. The measure ν_M is called the *Lévy measure* of M and A_M the *Gaussian variance*. Conversely, if $\gamma_M \in \mathbb{R}^d$, A_M is a symmetric non-negative definite $d \times d$ matrix, and ν_M is a Lévy measure, then there exists a Lévy process M, unique up to identity in law, such that (1) holds. The triplet (A_M, ν_M, γ_M) is called the *characteristic triplet* of the Lévy process M.

For Brownian motion $(X_t)_{t\geq 0}$ with $EX_t = \mu t$ and $\mathrm{Var}(X_t) = \sigma^2 t$, the characteristic triplet is $(\sigma^2, 0, \mu)$, and for a compound Poisson process with jump rate λ and jump-size distribution function F, the characteristic triplet is $(0, \lambda \mathrm{d}F(\cdot), \int_{[-1,1]} \lambda x \mathrm{d}F(x))$.

A Lévy process M with values in $\mathbb{R}^1$ is called a *subordinator* if it has increasing sample paths. This happens if and only if $A_M = 0$, $\nu_M((-\infty, 0)) = 0$, and $\int_0^1 x\, \nu_M(\mathrm{d}x) < \infty$. Examples of subordinators include compound Poisson processes with jump distribution concentrated on $(0, \infty)$, the Gamma process, and the inverse Gaussian process. The *Gamma process* with parameters $c, \lambda > 0$ is the Lévy process with characteristic triplet $(0, \nu_M, \int_0^1 c\mathrm{e}^{-\lambda x}\mathrm{d}x)$ and Lévy measure ν_M given by $\nu_M(\mathrm{d}x) = cx^{-1}\mathrm{e}^{-\lambda x}\mathbf{1}_{(0,\infty)}(x)\,\mathrm{d}x$. For the Gamma process, the distribution of M_t has Lebesgue density $x \mapsto (\Gamma(ct))^{-1}\lambda^{ct}x^{ct-1}\,\mathrm{e}^{-\lambda x}\mathbf{1}_{(0,\infty)}(x)$. The *inverse Gaussian* process with parameters $a, b > 0$ is defined to have characteristic triplet $A_M = 0$, Lévy measure $\nu_M(\mathrm{d}x) = (2\pi x^3)^{-1/2}a\mathrm{e}^{-xb^2/2}\mathbf{1}_{(0,\infty)}(x)\,\mathrm{d}x$, and $\gamma_M = 2ab^{-1}\int_0^b (2\pi)^{-1/2}\mathrm{e}^{-y^2/2}\,\mathrm{d}y$. For the inverse Gaussian process, the distribution of M_t has Lebesgue density $x \mapsto (2\pi x^3)^{-1/2}at\mathrm{e}^{-\frac{1}{2}(a^2t^2x^{-1}-2abt+b^2x)}$.

The *jump* of a Lévy process M at time t is defined as

$$\Delta X_t := X_t - X_{t-},$$

where X_{t-} denotes the left limit at $t > 0$ with the convention that $X_{0-} := 0$. Apart from Brownian motion with drift, every Lévy process has jumps. The Lévy measure $\nu_M(B)$ of a Borel set B describes the expected number of jumps of $(M_t)_{t\in[0,1]}$ with size in B, i.e.,

$$\nu_M(B) = E\sum_{0<s\leq 1} \mathbf{1}_B(\Delta M_s).$$

A Lévy process has only finitely many jumps in finite intervals if and only if the Lévy measure of the Lévy process is finite. Every one-dimensional Lévy process is a semimartingale (cf. Applebaum, 2004 or Protter, 2005), and its quadratic variation is given by $[M, M]_t = A_M t + \sum_{0<s\leq t} \Delta M_s^2$. We refer to Applebaum (2004) and Protter (2005) for further information regarding integration with respect to semimartingales (and in particular Lévy processes).

Finally, we mention that for $\kappa > 0$, a Lévy process $M = (M_t)_{t\geq 0}$ satisfies $E|M_1|^\kappa < \infty$ if and only if $E|M_t|^\kappa < \infty$ for all $t \geq 0$, which is further equivalent to $\int_{|x|\geq 1} |x|^\kappa\, \nu_M(\mathrm{d}x) < \infty$. In particular, for $\kappa = 2$ and $d = 1$, $\mathrm{Var}(M_t) = tA_M + \int_{\mathbb{R}} x^2\, \nu_M(\mathrm{d}x)$.

3. Lévy-driven CARMA(p, q) processes

If $(L_t)_{t\geq 0}$ is a Lévy process with values in $\mathbb{R}$, defined as in Section 2, it can be extended to a process with stationary independent increments, right-continuous sample paths with finite left limits, $L_0 = 0$ and index set $\mathbb{R}$, by defining $L_t = -M_{-t-}, t < 0$, where $(M_t)_{t\geq 0}$ is an independent version of $(L_t)_{t\geq 0}$. Assuming this extension has been made, we define an L-driven CARMA(p, q) process with real coefficients

$\{a_1, \ldots, a_p; b_1, \ldots, b_q\}$ and $p > q$ (see the work done by Brockwell, (2001)) as a strictly stationary solution of the suitably interpreted formal stochastic differential equation

$$a(D)V_t = b(D)DL_t, \quad t \in \mathbb{R}, \tag{2}$$

where D denotes differentiation with respect to t,

$$a(z) := z^p + a_1 z^{p-1} + \cdots + a_p,$$
$$b(z) := z^q + b_{q-1} z^{q-1} + \cdots + b_0.$$

Since DL_t does not exist in the usual sense, we interpret the differential equation (2) by means of its state space representation, consisting of the *observation* and *state* equations,

$$V_t = \mathbf{b}'\mathbf{X}_t \tag{3}$$

and

$$\mathrm{d}\mathbf{X}_t - \mathbf{A}\mathbf{X}_t \mathrm{d}t = \mathbf{e}\,\mathrm{d}L_t, \tag{4}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_p & -a_{p-1} & -a_{p-2} & \cdots & -a_1 \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-2} \\ b_{p-1} \end{bmatrix}, \tag{5}$$

$b_q := 1$ and $b_j := 0$ for $j > q$. Every solution of (4) satisfies the relations

$$\mathbf{X}_t = \mathrm{e}^{\mathbf{A}(t-s)}\mathbf{X}(s) + \int_s^t \mathrm{e}^{\mathbf{A}(t-u)}\mathbf{e}\,\mathrm{d}L_u, \quad t > s,$$

where the integral is a special case of integration with respect to a semimartingale.

Brockwell and Lindner (2009), Theorem 4.1, show that there is no loss of generality in assuming that $a(z)$ and $b(z)$ have no common factors and that, assuming L is nondeterministic, necessary and sufficient conditions for (3) and (4) to have a strictly stationary solution V are that $E\max(0, \log|L_1|)$ is finite and $a(z)$ is nonzero on the imaginary axis. In this case, the strictly stationary solution is unique and is given by

$$V_t = \int_{-\infty}^{\infty} g(t-u)\,\mathrm{d}L_u \tag{6}$$

with

$$g(t) = \left(\sum_{\lambda:\Re\lambda<0} \sum_{k=0}^{\mu(\lambda)-1} c_{\lambda k} t^k \mathrm{e}^{\lambda t} \mathbf{1}_{(0,\infty)}(t) - \sum_{\lambda:\Re\lambda>0} \sum_{k=0}^{\mu(\lambda)-1} c_{\lambda k} t^k \mathrm{e}^{\lambda t} \mathbf{1}_{(-\infty,0)}(t) \right), \tag{7}$$

where the sums are over the distinct zeroes λ of the polynomial $a(z)$ and $\mu(\lambda)$ denotes the multiplicity of λ. The sum $\sum_{k=0}^{\mu(\lambda)-1} c_{\lambda k} t^k \mathrm{e}^{\lambda t}$ is the residue of $z \mapsto \mathrm{e}^{zt} b(z)/a(z)$ at λ, i.e.,

$$\sum_{k=0}^{\mu(\lambda)-1} c_{\lambda k} t^k \mathrm{e}^{\lambda t} = \frac{1}{(\mu(\lambda)-1)!} \left[D_z^{\mu(\lambda)-1} \left((z-\lambda)^{\mu(\lambda)} \mathrm{e}^{zt} b(z)/a(z) \right) \right]_{z=\lambda},$$

and D_z denotes differentiation with respect to z. (For a zero λ with $\mu(\lambda) = 1$, the last sum reduces to $b(\lambda)e^{\lambda t}/a'(\lambda)$, where a' denotes the derivative of a.)

Remark 1 (Causality). The unique strictly stationary solution is causal if and only if $a(z)$ has no zeroes with positive real part, in which case the second sum in (7) disappears and g can be expressed as

$$g(t) = \begin{cases} \mathbf{b}' \mathrm{e}^{At} = \frac{1}{2\pi i} \int_\rho \frac{b(z)}{a(z)} \mathrm{e}^{tz} \mathrm{d}z, & \text{if } t > 0, \\ 0, & \text{if } t \le 0, \end{cases} \tag{8}$$

where the subscript ρ indicates integration anticlockwise around a simple closed contour encircling the zeroes of $a(z)$ and contained in the open left half of the complex plane. If the zeroes all have multiplicity 1, we obtain the very simple representation,

$$V_t = \sum_{j=1}^{p} \frac{b(\lambda_j)}{a'(\lambda_j)} \int_{-\infty}^{t} \mathrm{e}^{\lambda_j (t-u)} \mathrm{d}L_u. \tag{9}$$

From now on, we shall restrict attention to causal CARMA processes. The term stationary will be used to indicate strict (as opposed to weak or covariance) stationarity, except when explicitly stated otherwise. □

Example 1 (The stationary Ornstein–Uhlenbeck process). In the case when $p = 1$ (so that q is necessarily zero), V is the CARMA(1,0) process, also written as CAR(1) (continuous-time autoregression of order 1), and widely known as the stationary Ornstein–Uhlenbeck process. In this case, the dimension of the state vector $\mathbf{X}_t$ is 1, $b(z) = 1$, and $a(z) = z - a_1$ where, for causality, $a_1 = \lambda_1 < 0$. Provided $E \max(0, \log |L_1|) < \infty$, $V_t = \mathbf{X}_t$ is the unique stationary solution of the equation,

$$\mathrm{d}V_t - a_1 V_t \mathrm{d}t = \mathrm{d}L_t. \tag{10}$$

From (9), we immediately find that

$$V_t = \int_{-\infty}^{t} \mathrm{e}^{\lambda_1 (t-u)} \mathrm{d}L_u. \tag{11}$$

If L is a subordinator, i.e., a Lévy process with nondecreasing sample paths, then inspection of (11) shows that the process V is non-negative. The subordinator-driven CAR(1) process is thus a potential model for any non-negative process such as the stochastic volatility considered later in Section 4.

Example 2 (The CARMA(2,1) process). The CARMA(2,1) process with $a(z) = (z-\lambda_1)(z-\lambda_2)$, and $\lambda_1 \neq \lambda_2$, has the particularly simple structure,

$$V_t = \alpha_1 \int_{-\infty}^{t} \mathrm{e}^{\lambda_1(t-u)} \mathrm{d}L_u + \alpha_2 \int_{-\infty}^{t} \mathrm{e}^{\lambda_2(t-u)} \mathrm{d}L_u,$$

where

$$\alpha_i = \frac{b(\lambda_i)}{a'(\lambda_i)} = \frac{\lambda_i + b_0}{2\lambda_i + a_1}, \quad i = 1, 2.$$

The process is thus a sum of two dependent and possibly complex-valued CAR(1) processes. (Such a decomposition clearly extends to any CARMA(p,q) process for which $a(z)$ has distinct zeroes.) If L is a subordinator, then, as in Example 1, V_t is non-negative provided the kernel $g(t) = \alpha_1 \mathrm{e}^{\lambda_1 t} + \alpha_2 \mathrm{e}^{\lambda_2 t}$, $t \geq 0$, is non-negative. This is the case if and only if λ_1 and λ_2 are both real and $b_0 \geq \max(|\lambda_i|)$ (See the study by Brockwell and Davis (2001). More general conditions for non-negativity of a CARMA(p,q) kernel are given in the study by Tsai and Chan (2005)).

3.1. *Second-order properties when $EL_1^2 < \infty$*

If $EL_1^2 < \infty$, we define $\mu := EL_1$ and $\sigma^2 := \mathrm{Var}(L_1)$.

The causal CARMA process defined by (6) and (8) is then covariance stationary with mean $\mu b_0/a_p$ and autocovariance function, which can be calculated as follows. From (8), noting that $g(t) = 0$ for $t < 0$, we see that the Fourier transform of g is

$$\tilde{g}(\omega) := \int_{\mathbb{R}} g(t) \mathrm{e}^{i\omega t} \mathrm{d}t = -\frac{1}{2\pi i} \int_{\rho} \frac{b(z)}{a(z)} \frac{1}{z + i\omega} \mathrm{d}z = \frac{b(-i\omega)}{a(-i\omega)}, \quad \omega \in \mathbb{R}.$$

Since the autocovariance function $\gamma_V(\cdot)$ is the convolution of $\sigma g(\cdot)$ and $\sigma g(-\cdot)$, its Fourier transform is given by

$$\tilde{\gamma}_V(\omega) = \sigma^2 \tilde{g}(\omega)\tilde{g}(-\omega) = \sigma^2 \left| \frac{b(i\omega)}{a(i\omega)} \right|^2, \quad \omega \in \mathbb{R}.$$

The spectral density of V is the inverse Fourier transform of γ_V. Thus,

$$f_V(\omega) = \frac{1}{2\pi} \int_{\mathbb{R}} \mathrm{e}^{-i\omega h} \gamma_V(h) \mathrm{d}h = \frac{1}{2\pi} \tilde{\gamma}_V(-\omega) = \frac{\sigma^2}{2\pi} \left| \frac{b(i\omega)}{a(i\omega)} \right|^2, \quad \omega \in \mathbb{R}.$$

Substituting this expression into the relation

$$\gamma_V(h) = \int_{\mathbb{R}} e^{i\omega h} f_V(\omega) d\omega, \quad h \in \mathbb{R}$$

and changing the variable of integration from ω to $z = i\omega$ gives

$$\gamma_V(h) = \frac{\sigma^2}{2\pi i} \int_{\rho} \frac{b(z)b(-z)}{a(z)a(-z)} e^{|h|z} dz = \sigma^2 \sum_{\lambda} Res_{z=\lambda} \left(e^{z|h|} b(z)b(-z)/(a(z)a(-z)) \right),$$

where the sum is again over the distinct zeroes, λ of $a(z)$. This gives the general expression

$$\gamma_V(h) = \sigma^2 \sum_{\lambda} \frac{1}{(\mu(\lambda)-1)!} \left[D_z^{\mu(\lambda)-1} \frac{(z-\lambda)^m e^{z|h|} b(z)b(-z)}{a(z)a(-z)} \right]_{z=\lambda}, \tag{12}$$

where $\mu(\lambda)$ is the multiplicity of λ. In the case when the roots are distinct, Eq. (12) simplifies to

$$\gamma_V(h) = \sigma^2 \sum_{\lambda: a(\lambda)=0} \frac{e^{\lambda|h|} b(\lambda)b(-\lambda)}{a'(\lambda)a(-\lambda)}. \tag{13}$$

Example 3 (The second-order CAR(1) process). If $EL_1^2 < \infty$, then, by (13), the CAR(1) process defined in Example 1 has the autocovariance function

$$\gamma_V(h) = \frac{\sigma^2}{2|\lambda_1|} e^{\lambda_1|h|} \tag{14}$$

and autocorrelation function $\rho_V(h) = e^{\lambda_1|h|}$.

Example 4 (The second-order CARMA(2,1) process). If $EL_1^2 < \infty$, then, by (13), the CARMA(2,1) process defined in Example 2 with $\lambda_1 \neq \lambda_2$ has the autocovariance function,

$$\gamma_V(h) = \frac{\sigma^2}{2\lambda_1\lambda_2(\lambda_1^2 - \lambda_2^2)} \left[\lambda_2(b_0^2 - \lambda_1^2) e^{\lambda_1|h|} - \lambda_1(b_0^2 - \lambda_2^2) e^{\lambda_2|h|} \right].$$

This is a much broader class of functions than those in Example 3, allowing the approximation of a wider class of sample autocovariances than is possible when attention is restricted to CAR(1) models.

4. A continuous-time stochastic volatility model

Let λ be strictly negative and let L be a subordinator. Then the spot volatility process V in the stochastic volatility model of Barndorff-Nielsen and Shephard (2001) is defined, apart from a change of time scale, as the strictly stationary solution of the equation

$$dV_t = \lambda V_t \, dt + dL_t, \tag{15}$$

i.e., as a subordinator-driven CAR(1) process with driving Lévy process L and coefficient λ. Then, by (11), V_t is positive for all $t \in \mathbb{R}$. If G_t denotes the logarithm of the asset price at time t, then the process $(G_t)_{t\geq 0}$ is assumed to satisfy the stochastic differential equation,

$$\mathrm{d}G_t = (m + bV_t)\,\mathrm{d}t + \sqrt{V_t}\,\mathrm{d}W_t, \tag{16}$$

where m and b are constants, and $(W_t)_{t\geq 0}$ is a standard Brownian motion, independent of L.

NOTATION 1. The term volatility is sometimes used to refer to V_t and sometimes to $\sqrt{V_t}$. We shall refer to V_t as the (spot) volatility at time t and to integrals of V_t over time intervals as integrated volatility. □

If $EL_1^2 < \infty$ and $\mathrm{Var}(L_1) = \sigma^2$, the autocovariance function of V is, by (14), the exponentially decaying function,

$$\mathrm{Cov}(V_{t+h}, V_t) = \sigma^2 \mathrm{e}^{\lambda|h|}/(2|\lambda|).$$

If additionally $m = b = 0$, then (see the work done by Barndorff-Nielsen and Shephard (2001, Section 4)) nonoverlapping increments of G of length $r > 0$ are uncorrelated, i.e.,

$$\mathrm{Cov}(G_t - G_{t-r}, G_{t+h} - G_{t+h-r}) = 0, \quad t, h \geq r,$$

while, if $EL_1^4 < \infty$, the squared increments are correlated with the autocovariance function,

$$\mathrm{Cov}((G_t - G_{t-r})^2, (G_{t+h} - G_{t+h+r})^2) = C_r \mathrm{e}^{-\lambda h}$$

for strictly positive integer multiples h of $r > 0$, where $C_r > 0$ is some constant. The process $((G_{rh} - G_{r(h-1)})^2)_{h\in\mathbb{N}}$ thus has the autocovariance structure of an ARMA(1,1) process. The fact that the increments of the log-price process are uncorrelated while its squares are not is one of the important *stylized features* of financial time series. The tail behavior of the squared volatility process depends on the tail behavior of the driving Lévy process. In particular, it can be seen that V_t has Pareto tails, i.e., that $P(V_t > x)$ behaves asymptotically as a constant times $x^{-\alpha}$ for some $\alpha > 0$ as $x \to \infty$ if and only if L_1 has Pareto tails with the same index α (see the study by Fasen et al. (2006); the converse follows from the monotone density theorem for regularly varying functions; see, e.g., Theorem 1.7.2 in the study by Bingham et al., 1987).

5. Integrated CARMA processes and spot volatility modeling

In the stochastic volatility model (15) and (16), the spot volatility V_t is represented by a stationary Lévy-driven Ornstein–Uhlenbeck process. This has the shortcoming

that its autocorrelation function is necessarily a decreasing exponential function. Spot volatility is not an observable quantity, however, the *integrated volatility sequence*

$$I_n^{\Delta} = \int_{(n-1)\Delta}^{n\Delta} V_t \, \mathrm{d}t, \quad n = 1, 2, \ldots \tag{17}$$

over successive periods of length Δ can be well estimated in the context of the model (15) and (16) by the so-called *realized volatility sequence*,

$$R_n = \sum_{j=1}^{k} d_{n,j}^2, \tag{18}$$

where

$$d_{n,j} = (G_{(n-1+j/k)\Delta} - G_{(n-1+(j-1)/k)\Delta})^2,$$

and k is large. Typically Δ denotes a single trading day and k is such that Δ/k is a 5 min interval. An excellent discussion of realized volatility can be found in the article by Andersen and Benzoni (2009). Figure 1 shows the realized daily volatility (kindly supplied by Viktor Todorov) of the Deutsche Mark/US dollar exchange rate from December 1, 1986, through June 30, 1999. (See the study by Andersen et al. (2001) for a discussion of the series on which this realized volatility was based.)

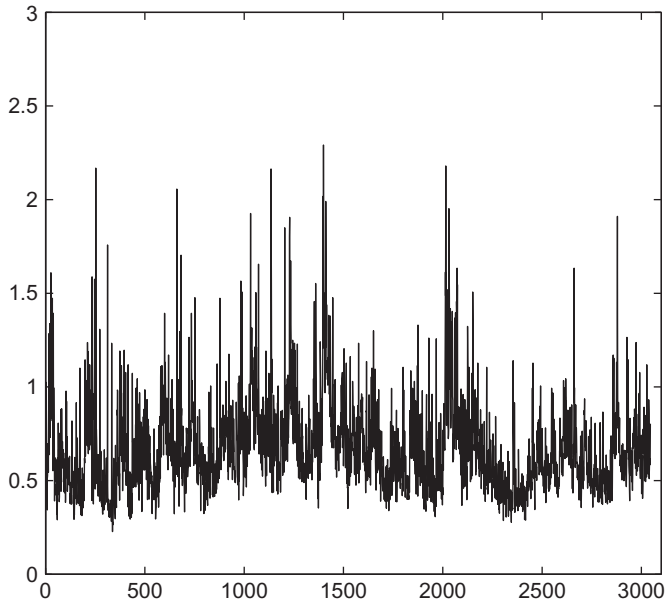

Fig. 1. The realized daily volatility of the DM/US$ exchange rate, December 1, 1986, through June 30, 1999.

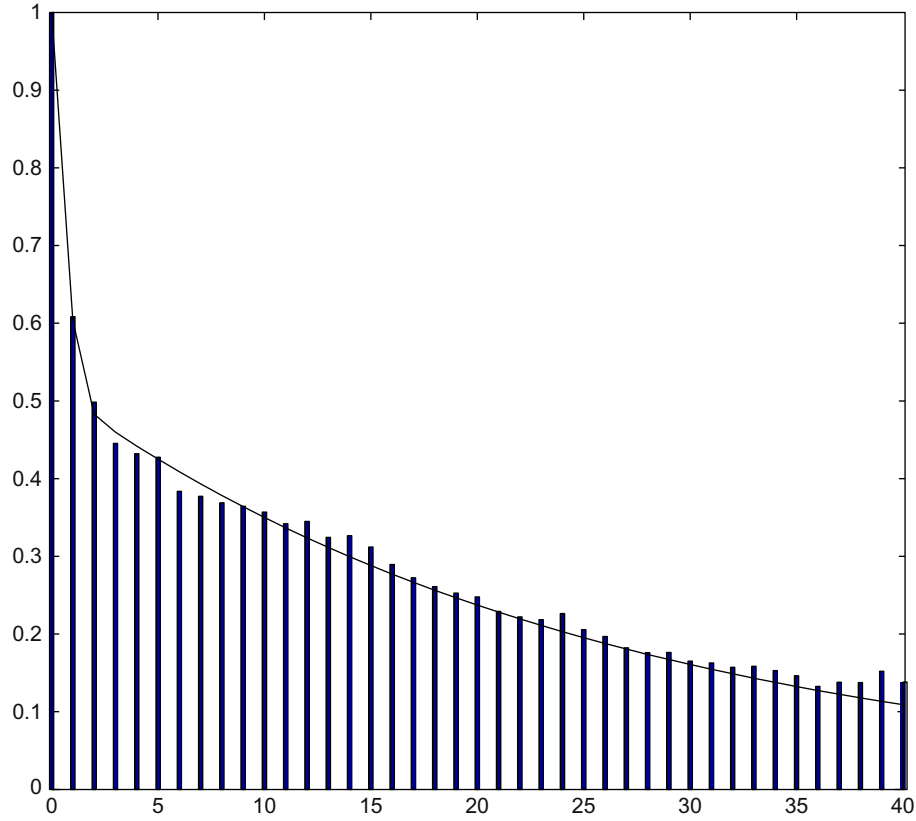

Fig. 2. The vertical bars represent the sample autocorrelation function of the daily realized volatility of the DM/US$ exchange rate shown in Fig. 1. The line graph is the autocorrelation function of the integrated volatility corresponding to a CARMA(2,1) model for spot volatility, estimated as described in Example 5.

The sample autocorrelation function of the series is shown in Fig. 2. Under the CAR(1) model (15) for the spot volatility V_t, it has been shown in the study by Barndorff-Nielsen and Shephard (2001) that the daily integrated volatility is an ARMA(1,1) process so that its autocorrelation function at lags greater than zero is a decreasing exponential function. It is clear from Fig. 2 that a better fit should be achievable by modeling V as a higher order CARMA process. Todorov and Tauchen (2006), Todorov (2010), and Brockwell et al. (2011) fitted CARMA(2,1) models to the *daily realized volatility*. In this section, we take a different point of view, the goal being to replace (15) by a CARMA model for the *s*pot volatility V in such a way that the corresponding integrated volatility sequence (17) adequately reflects the properties of the observed realized volatility sequence (18).

If $EL_1^2 < \infty$ and V is the L-driven CARMA(p,q) process with autoregressive and moving-average polynomials $a(z)$ and $b(z)$, respectively, the study by Brockwell and Lindner (2012) show that the integrated volatility sequence I^{Δ} is a weak stationary solution of the difference equations,

$$\phi(B)I_n^{\Delta} = \theta(B)\epsilon_n,$$

where $(\epsilon_n)_{n\in\mathbb{Z}}$ is an uncorrelated constant variance sequence, B is the backward shift operator (i.e., $B^j Y_n := Y_{n-j}$, for all j and $n \in \mathbb{Z}$) and $\phi(z)$ is the polynomial,

$$\phi(z) := \prod_{\lambda}(1 - \mathrm{e}^{\lambda\Delta}z)^{\mu(\lambda)},$$

where the product is over the distinct zeroes λ of $a(z)$, and $\mu(\lambda)$ denotes the multiplicity of λ. The polynomial $\theta(z)$ has the form,

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_p z^p,$$

with coefficients $\theta_1, \ldots, \theta_p$ that can be determined from $a(z)$ and $b(z)$ and chosen to have no zeroes in the interior of the unit disc. For any $a(z)$ and $b(z)$, the corresponding ARMA polynomials $\phi(z)$ and $\theta(z)$ for the ARMA process I^Δ can therefore be determined and hence the minimum mean-squared-error one-step linear predictors of the sequence I^Δ. Numerical minimization of the sum of squares of these one-step errors with respect to the coefficients of the polynomials $a(z)$ and $b(z)$ gives least squares estimates of the CARMA coefficients for the spot volatility process V.

Example 5. To illustrate the procedure, we consider the daily realized volatility in Fig. 1. It is clear that a good match between the sample autocorrelation function in Fig. 2 for lags greater than zero and a single exponential function (as would be derived from an Ornstein–Uhlenbeck model for spot volatility) is not possible. We therefore try a CARMA(2,1) model for the spot volatility. Measuring the spot volatility in units of volatility per day, the realized volatility series corresponds to volatility integrated over time intervals of length 1, i.e., I^Δ with $\Delta = 1$.

A simple initial guess at appropriate values of the coefficients can be obtained by attempting to match the autocorrelation function of I^Δ with the sample autocorrelation of the realized volatility V^Δ at selected lags. If, for example, we minimize the sum of squared differences at lags 1, 2, 10, 20, and 40, we obtain the preliminary spot volatility model,

$$(D^2 + 3.09054D + .10983)V_t = (.23302 + D)DL_t,$$

with corresponding $\lambda_1 = -0.035956$ and $\lambda_2 = -3.05458$.

Using these coefficients as initial values, the numerical minimization of the prediction sum of squares leads to the least-squares model,

$$(D^2 + 3.07141D + .11793)V_t = (.23938 + D)DL_t, \tag{19}$$

with corresponding $\lambda_1 = -0.038890$ and $\lambda_2 = -3.02152$. The autocorrelation function of the daily integrated volatility corresponding to this model is plotted as the line graph in Fig. 2.

It remains to identify a subordinator L that yields daily integrated volatilities compatible with the realized daily volatility series shown in Fig. 1. This was done by trying several subordinators, each with $EL_1 = 0.3291$ and $\mathrm{Var}L_1 = 0.3954$ to match the mean and variance of the realized volatility series, simulating sample paths of

the corresponding CARMA(2,1) processes defined by (19), integrating the sample paths over successive days, and comparing the empirical cumulative distribution functions and kernel density estimates of the realized volatility series with those of the integrated volatilities calculated from the models. The results are shown in Fig. 3.

The top graphs were generated by simulating the CARMA(2,1) process (19) driven by a compound Poisson subordinator with exponentially distributed jumps. The mean jump rate of the process was 0.5478 and the mean jump size was 0.6008. The simulation of the CARMA process is greatly simplified by the decomposition in Example 3, which reduces the simulation to that of two Ornstein–Uhlenbeck processes with the same driving subordinator. In fact from the simulated jump times and jump sizes, the complete sample path can be constructed and the daily integrals can be easily computed. The same is true for compound-Poisson-driven CARMA processes of any order as long as the zeroes of $a(z)$ are distinct.

The middle graphs are derived from the spot volatility process (19) with inverse Gaussian subordinator having $EL(1) = 0.3291$ and $\mathrm{Var}L(1) = 0.3954$. Simulation

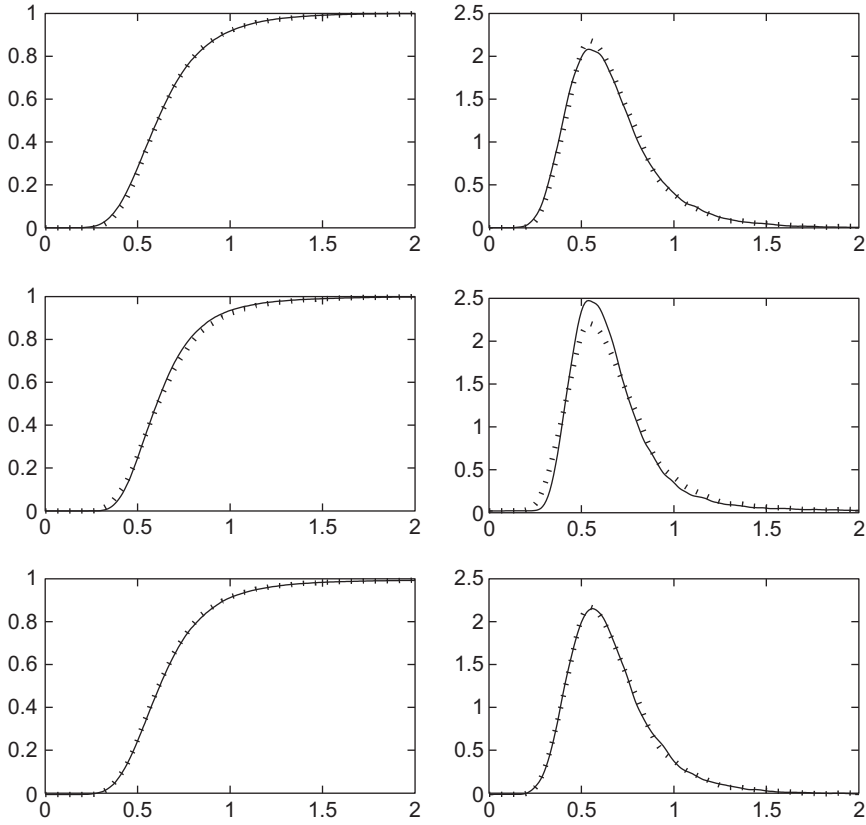

Fig. 3. The empirical cdf (left) and kernel density estimate (right) of the daily realized volatility of the DM/US$ exchange rate are shown as dotted lines. The solid lines are the corresponding graphs for daily integrated volatility of three subordinator-driven CARMA(2,1) spot volatility processes. For details see Example 5.

in this case was carried out by using an Euler approximation to generate values of the spot volatility at intervals of 0.01 days and integrating numerically to get 40,000 daily integrated volatility values.

The bottom graphs were derived in the same way, using a gamma subordinator with $EL(1) = 0.3291$ and $\mathrm{Var}L(1) = 0.3954$. The empirical cdf and kernel density estimates were again based on 40,000 daily integrated volatility values.

The reasonable fits by all three subordinators suggest that the distribution of daily integrated volatilities is rather insensitive to the distribution of L_1; however, in the case of the gamma subordinator, the empirical and simulated distributions are virtually indistinguishable.

6. Generalized Ornstein–Uhlenbeck processes

The Lévy-driven CARMA processes can be regarded as a higher order generalization of the Ornstein–Uhlenbeck process. From (10), the Lévy-driven Ornstein–Uhlenbeck process satisfies the equation,

$$\mathrm{d}V_t = \lambda_1 V_{t-}\mathrm{d}t + \mathrm{d}L_t, \quad t \geq 0.$$

Another way in which to extend the class of Ornstein–Uhlenbeck processes is to replace the deterministic function $t \mapsto \lambda_1 t$ in this equation by a second Lévy process $(U_t)_{t\geq 0}$. This leads to the stochastic differential equation

$$\mathrm{d}V_t = V_{t-}\,\mathrm{d}U_t + \mathrm{d}L_t, \quad t \geq 0, \tag{20}$$

or equivalently,

$$V_t = V_0 + \int_0^t V_{s-}\,\mathrm{d}U_s + L_t, \quad t \geq 0,$$

where V_0 is a starting random variable and (U, L) is a bivariate Lévy process. We shall always assume that neither U nor L is the zero process. This stochastic differential equation was considered and solved by Yoeurp and Yor (1977) (cf. Protter, 2005, Exercise V.27), see also the study by Behme et al. (2011). In particular, it is shown that if U has no jumps of size -1, i.e., if $\nu_U(\{-1\}) = 0$, then the unique solution to (20) is given by

$$V_t = \mathcal{E}(U)_t \left(V_0 + \int_0^t [\mathcal{E}(U)_{s-}]^{-1}\,\mathrm{d}\eta_s \right), \quad t \geq 0. \tag{21}$$

Here, $\mathcal{E}(U)$ denotes the stochastic exponential of U, given by

$$\mathcal{E}(U)_t = \mathrm{e}^{U_t - t\sigma_U^2/2} \prod_{0<s\leq t} (1 + \Delta U_s)\mathrm{e}^{-\Delta U_s}, \quad t \geq 0,$$

and η is given by

$$\eta_t = L_t - \sum_{0<s\le t} \frac{\Delta U_s \Delta L_s}{1+\Delta U_s} - t\sigma_{U,L}, \quad t \ge 0,$$

where σ_U^2 and $(\sigma_{U,L})$ denote the $(1,1)$ and $(1,2)$ elements of the Gaussian variance $A_{(U,L)}$ of (U,L), respectively.

When U has no jumps of size less than or equal to -1, i.e., when $\nu_U((-\infty,-1]) = 0$, then $\mathcal{E}(U)_t$ is strictly positive and we can define

$$\xi_t := -\log \mathcal{E}(U)_t = -U_t + \sigma_U^2 t/2 + \sum_{0<s\le t} (\Delta U_s - \log(1+\Delta U_s)), \quad t \ge 0.$$

Then $(\xi,\eta) = (\xi_t,\eta_t)_{t\ge 0}$ is again a bivariate Lévy process, and the process in (21) can be written as

$$V_t = \mathrm{e}^{-\xi_t}\left(V_0 + \int_0^t \mathrm{e}^{\xi_{s-}}\, d\eta_s\right), \quad t \ge 0. \tag{22}$$

If V_0 is additionally independent of (ξ,η) (equivalently, of (U,L)), then the process given by (22) is called a *generalized Ornstein–Uhlenbeck process*, driven by (ξ,η). This terminology is due to de Haan and Karandikar (1989) and Carmona et al. (1997), who studied various properties of these processes. The process (U,L) can be recovered from (ξ,η) by

$$\begin{pmatrix} U_t \\ L_t \end{pmatrix} = \begin{pmatrix} -\xi_t + \sum_{0<s\le t}\left(\mathrm{e}^{-\Delta\xi_s} - 1 + \Delta\xi_s\right) + t\,\sigma_\xi^2/2 \\ \eta_t + \sum_{0<s\le t}(\mathrm{e}^{-\Delta\xi_s} - 1)\Delta\eta_s - t\,\sigma_{\xi,\eta} \end{pmatrix}, \quad t \ge 0.$$

Obviously, if U and L are independent, then so are ξ and η (and conversely), in which case $\eta_t = L_t$.

It is clear that if $\xi_t = -U_t = -\lambda_1 t$, $t \ge 0$, then a generalized Ornstein–Uhlenbeck process reduces to the Lévy-driven Ornstein–Uhlenbeck process defined by (10).

We have already seen that generalized Ornstein–Uhlenbeck processes arise as Lévy-driven Ornstein–Uhlenbeck processes in the stochastic volatility model of Barndorff-Nielsen and Shephard (2001). In Section 7, we shall see that generalized Ornstein–Uhlenbeck processes also arise as volatility processes of continuous-time GARCH(1,1) processes, when ξ is deterministic and η random in contrast to the situation of the Lévy-driven Ornstein–Uhlenbeck process. In general, since generalized Ornstein–Uhlenbeck processes are the natural continuous time analogues of AR(1) processes with random i.i.d. coefficients (cf. the study by de Haan and Karandikar, 1989), a non-negative generalized Ornstein–Uhlenbeck process may serve as a stochastic volatility model and hence has potential applications in finance, even when not restricted to the continuous-time GARCH situations. The generalized Ornstein–Uhlenbeck process is non-negative if $V_0 \ge 0$ and η is a subordinator.

Another branch of finance in which generalized Ornstein–Uhlenbeck processes make an appearance is insurance mathematics, specifically in the risk model of Paulsen

(1993). Here, V_t denotes the capital of an insurance company, L_t describes the premium minus the claim process, and U describes the behavior of a financial market in which the capital of the insurance company is invested. See the study by Paulsen (1993) for details.

For using a generalized Ornstein–Uhlenbeck process as a volatility model, it is interesting to know for which bivariate Lévy processes (ξ, η) there exists a starting random variable V_0, independent of (ξ, η), such that the corresponding generalized Ornstein–Uhlenbeck process becomes strictly stationary. A complete characterization of this was obtained in the study by Lindner and Maller (2005). Accordingly, a strictly stationary solution exists if and only if there is $k \in \mathbb{R} \setminus \{0\}$ such that $e^{\xi} = \mathcal{E}(\eta/k)$, in which case $V_t = k$ for all $t \geq 0$, or if the integral $\int_0^t e^{-\xi_{s-}} \, dL_s$ converges almost surely as $t \to \infty$, in which case the marginal stationary distribution is given by the distribution of

$$\int_0^\infty e^{-\xi_{s-}} \, dL_s. \tag{23}$$

A necessary and sufficient condition for the integral in (23) to converge almost surely absolutely has been given in the study by Erickson and Maller (2004). A sufficient condition for the convergence of the integral is that $E \log^+ |L_1| < \infty$ and that ξ_t converges almost surely to $+\infty$, the latter being implied by $E\xi_1 > 0$ (cf. the study by de Haan and Karandikar, 1989 and Lindner and Maller, 2005). An extension of the characterization of stationary solutions of generalized Ornstein–Uhlenbeck processes to solutions of (20), when U is allowed to have jumps of size less than or equal to -1 and V_0 is allowed also to be dependent of (U, L), hence allowing also noncausal solutions, has been given in the study by Behme et al. (2011).

The autocorrelation structure of a generalized Ornstein–Uhlenbeck is always of exponential form. More precisely, if U and L are such that $\nu_U((-\infty, -1]) = 0$, $EU_1^2, EL_1^2 < \infty$ and $E\mathcal{E}(U)_1^2 = Ee^{-2\xi_1} < 1$, then $EU_1 < 0$ and a stationary version with finite second moment of the generalized Ornstein–Uhlenbeck process exists, the mean of which is given by $EV_0 = -(EU_1)^{-1} EL_1$ and the autocovariance function by

$$\mathrm{Cov}(V_t, V_{t+h}) = \frac{E(U_1 EL_1 - L_1 EU_1)^2}{(EU_1)^2 |2EU_1 + \mathrm{Var}U_1|} e^{hEU_1}, \quad t, h \geq 0;$$

see also the study by Behme (2011a,b).

Another important feature of generalized Ornstein–Uhlenbeck processes is that they allow the stationary solution to have Pareto tails for a wide variety of situations, even if η does not have heavy tails. This follows from the results of Kesten (1973) and Goldie (1991) on the tail behavior of solutions of random recurrence equations; see the study by Behme (2011a,b) and Lindner and Maller (2005) for details. Finally, we remark that multivariate extensions of generalized Ornstein–Uhlenbeck processes have been recently obtained in the study by Behme (2011b); see also the study by Behme and Lindner (2012).

7. Continuous-time GARCH processes

Among the most prominent discrete time models for financial time series are the ARCH and GARCH processes of Engle (1982) and Bollerslev (1986). Given an i.i.d. sequence $(\varepsilon_n)_{n\in\mathbb{N}_0}$ and constants $\beta > 0$, $\lambda_1, \dots, \lambda_q \geq 0$ and $\delta_1, \dots, \delta_p \geq 0$ with $q \in \mathbb{N}$ and $p \in \mathbb{N}_0$ and $\lambda_q > 0$, a GARCH(q, p) process $(Y_n)_{n\in\mathbb{N}_0}$ with volatility process $(V_n)_{n\in\mathbb{N}_0}$ is given by

$$Y_n = \sqrt{V_n}\,\varepsilon_n, \quad n \in \mathbb{N}_0, \tag{24}$$

$$V_n = \beta + \sum_{i=1}^{q} \lambda_i Y_{n-i}^2 + \sum_{j=1}^{p} \delta_j V_{n-j}, \quad n \geq \max\{p, q\}, \tag{25}$$

with V_n independent of $(\varepsilon_{n+h})_{h\in\mathbb{N}_0}$ and non-negative for every $n \in \mathbb{N}_0$. For $p = 0$, the process is called an ARCH(q) process.

Continuous-time diffusion limits have been obtained in the study by Nelson (1990) for the GARCH(1,1) process and in the study by Duan (1997) for the GARCH(q, p) process. By considering GARCH processes on fine grids $h\mathbb{N}_0$ and rescaling the parameters appropriately as $h \downarrow 0$, Nelson obtained the following diffusion limit $(G_t, V_t)_{t\geq 0}$ given by

$$\mathrm{d}G_t = \sqrt{V_t}\,\mathrm{d}B_t^{(1)}, \quad t \geq 0, \tag{26}$$

$$\mathrm{d}V_t = (\omega - \theta V_t)\,\mathrm{d}t + \alpha V_t\,\mathrm{d}B_t^{(2)}, \quad t \geq 0, \tag{27}$$

where $B^{(1)}$ and $B^{(2)}$ are two independent Brownian motions and $\theta \in \mathbb{R}$, $\omega \geq 0$, and $\alpha > 0$ are parameters. In particular, the volatility process determined by (27) is a generalized Ornstein–Uhlenbeck process driven by $(\xi_t, \eta_t) = (-\alpha B_t^{(2)} + (\theta + \alpha^2/2)t, \omega t)$. It should be observed that the diffusion limit of Nelson has two independent sources of randomness, namely $B^{(1)}$ and $B^{(2)}$, while the GARCH(1,1) process defined by (24) and (25) is driven by a single noise process $(\varepsilon_n)_{n\in\mathbb{N}_0}$. This motivated Klüppelberg et al. (2004) to construct a continuous-time GARCH(1,1) process driven by a single Lévy process, called COGARCH(1,1). Given a driving Lévy process $M = (M_t)_{t\geq 0}$ with nonzero Lévy measure, independent of a starting random variable $V_0 \geq 0$, and constants $\beta, \delta > 0$, and $\lambda \geq 0$, they define the *COGARCH(1,1) process* $(G_t)_{t\geq 0}$ with *volatility process* $(V_t)_{t\geq 0}$ by

$$G_0 = 0, \quad \mathrm{d}G_t = \sqrt{V_{t-}}\,\mathrm{d}M_t, \quad t \geq 0,$$

where

$$V_t = \left(\beta \int_0^t \mathrm{e}^{\xi_{s-}}\,\mathrm{d}s + V_0\right)\mathrm{e}^{-\xi_t}, \quad t \geq 0,$$

and $\xi = (\xi_t)_{t\geq 0}$ is defined by

$$\xi_t := -t\log\delta - \sum_{0<s\leq t} \log(1 + \lambda\delta^{-1}(\Delta M_s)^2), \quad t \geq 0.$$

Then ξ is again a Lévy process and V is a generalized Ornstein–Uhlenbeck process, driven by the bivariate Lévy process $(\xi_t, \beta t)_{t\geq 0}$. The corresponding processes (U, L) in the differential equation (20) are given by $U_t = t\log\delta + \lambda\delta^{-1}\sum_{0<s\leq t}(\Delta M_s)^2$ and $L_t = \beta t$, i.e.,

$$\mathrm{d}V_t = V_{t-}\mathrm{d}(t\log\delta + \lambda\delta^{-1}[M,M]_t^{(d)}) + \beta\mathrm{d}t, \quad t \geq 0,$$

where $[M,M]_t^{(d)} = \sum_{0<s\leq t}(\Delta M_s)^2$ denotes the discrete part of the quadratic variation of M. A multivariate extension of the COGARCH(1,1) process has been obtained in the study by Stelzer (2010).

It has been shown in the study by Klüppelberg et al. (2004) that a stationary volatility process of the COGARCH(1,1) equations exists if and only if

$$\int_{\mathbb{R}} \log(1+\lambda\delta^{-1}x^2)\,\nu_M(\mathrm{d}x) < -\log\delta, \tag{28}$$

which in particular requires M to have finite log-moment and $\delta < 1$. The second moment structure of V_t can be obtained from those of generalized Ornstein–Uhlenbeck processes. More precisely, under the condition (28) and defining

$$\Psi_\xi(\kappa) := \log E\mathrm{e}^{-\kappa\xi_1} = \kappa\log\delta + \int_{\mathbb{R}} \left((1+\lambda\delta^{-1}y^2)^\kappa - 1\right)\nu_M(dy) \in (-\infty,\infty]$$

for $\kappa > 0$, the stationary version has the property for $k \in \mathbb{N}$ that $EV_0^k < \infty$ if and only if $EM_1^{2k} < \infty$ and $\Psi_\xi(k) < 0$, in which case $\Psi_\xi(l) < 0$ for all $l \in \{1,\ldots,k\}$ and

$$EV_0^k = k!\,\beta^k \prod_{l=1}^{k}(-\Psi_\xi(l))^{-1}.$$

Further, if $EM_1^4 < \infty$ and $\Psi_\xi(2) < 0$, then

$$\mathrm{Cov}(V_t, V_{t+h}) = \beta^2\left(2\Psi_\xi^{-1}(1)\Psi_\xi^{-1}(2) - \Psi_\xi^{-2}(1)\right)\mathrm{e}^{-h|\Psi_\xi(1)|}, \quad t,h \geq 0.$$

A detailed proof is given in the study by Klüppelberg et al. (2004).

As is the case for the volatility model of Barndorff-Nielsen and Shephard considered in Section 4, under certain assumptions, non-overlapping increments of the stationary COGARCH(1,1) process G are uncorrelated, while the autocovariance function of $((G_{rh} - G_{r(h-1)})^2)_{h\in\mathbb{N}}$ is that of an ARMA(1,1) process for any $r > 0$. More precisely, restricting to $r = 1$ for simplicity, if the driving Lévy process $M = (M_t)_{t\geq 0}$ satisfies

$$EM_1 = 0, \quad \mathrm{Var}(M_1) = 1, \quad EM_1^4 < \infty, \quad \int_{\mathbb{R}} x^3\,\nu_M(\mathrm{d}x) = 0$$

and if

$$\Psi_\xi(2) = 2\log\delta + \int_{\mathbb{R}} \left(\lambda^2\delta^{-2}y^4 + 2\lambda\delta^{-1}y^2\right)\nu_\xi(dy) < 0,$$

then the increment process $(Y_n)_{n\in\mathbb{N}}$ with $Y_n = G_n - G_{n-1}$ satisfies $EY_1^4 < \infty$ and

$$EY_1 = 0, \quad \mu := E(Y_1^2) = \frac{\beta}{|\Psi_\xi(1)|} \quad \text{and} \quad \operatorname{Cov}(Y_t, Y_{t+h}) = 0, \quad t, h \in \mathbb{N}.$$

Denoting

$$\varphi := \lambda\delta^{-1} \quad \text{and} \quad \tau := -\log\delta,$$

the autocorrelation function ρ of Y satisfies

$$\rho(h) = k\mathrm{e}^{-hp}, \quad t, h \in \mathbb{N}, \tag{29}$$

where

$$p := |\Psi_\xi(1)|$$

and

$$k := \frac{\beta^2}{p^3\gamma(0)}(2\tau\varphi^{-1} + 2A_M - 1)\left(2|\Psi_\xi^{-1}(2)| - p^{-1}\right)\left(1 - \mathrm{e}^{-p}\right)(\mathrm{e}^p - 1).$$

An explicit expression for $\operatorname{Var}(Y_0)$ can also be obtained. Based on these expressions, Haug et al. (2007) consider a generalized method of moment estimator for the parameters of the COGARCH(1,1) process, by replacing $E(Y_1^2)$, $\operatorname{Var}(Y_0)$, and $\log\rho(h)$ by their empirical counterparts and doing a regression for p and k in (29). Assuming that the Gaussian variance A_M is known (e.g., $A_M = 0$), solving the equations obtained for $\mu, \operatorname{Var}(Y_0), p, k$ in β, φ, τ, and plugging the obtained estimators $\widehat{\mu}$, $\widehat{\operatorname{Var}(Y_0)}$, $\widehat{p}$, and $\widehat{k}$ into these equations gives generalized method of moment estimators $(\widehat{\beta}, \widehat{\varphi}, \widehat{\tau})$ for the parameters (β, φ, τ) and hence for (β, δ, λ) based on observations $G_0, G_1, G_2, \ldots, G_n$. Details can be found in the study by Haug et al. (2007). There it is also shown that the estimator is strongly consistent and under further moment assumptions, which require in particular a finite 8th moment of Y, that the estimator is asymptotically normal.

Other estimation methods for the COGARCH(1,1) include the pseudo-maximum likelihood estimator of Maller et al. (2008) and the Markov Chain Monte Carlo estimator of Müller (2010). Maller et al. (2008) also fit the COGARCH(1,1) model to the ASX200 index of the Australian Stock exchange.

Finally, let us introduce the COGARCH(q, p) processes of Brockwell et al. (2006). From (24) and (25), we see that the volatility (V_n) of a GARCH(q, p) process can be regarded as a "self-exciting" ARMA$(p, q-1)$ process driven by $(V_{n-1}\varepsilon_{n-1}^2)$ together with the "mean correction" β. This motivates the definition of the volatility process $(V_t)_{t\geq 0}$ of a continuous-time GARCH(q, p) process as a "self-exciting mean corrected" CARMA$(p, q-1)$ process driven by an appropriate noise term. Since in discrete time, the driving noise is defined through the increments of the process $(\sum_{i=0}^{n-1} V_i\varepsilon_i^2)_{n\in\mathbb{N}}$, in continuous time this suggests the use of

$$R_t = \sum_{0<s\leq t} V_{s-}(\Delta M_s)^2 = \int_0^t V_{s-}d[M, M]_s^{(d)}, \quad t \geq 0,$$

as driving noise for the CARMA equations. More precisely, let $M = (M_t)_{t\geq 0}$ be a Lévy process with nonzero Lévy measure. With $p, q \in \mathbb{N}$ such that $q \leq p$, $a_1, \ldots, a_p, b_0, \ldots, b_{p-1} \in \mathbb{R}, \beta > 0, a_p \neq 0, b_{q-1} \neq 0,$ and $b_q = \cdots = b_{p-1} = 0,$ define the $p \times p$ matrix $\mathbf{A}$ and the vectors $\mathbf{b}, \mathbf{e} \in \mathbb{C}^p$ as in (5). Define the volatility process $(V_t)_{t\geq 0}$ with parameters $\mathbf{A}$, $\mathbf{b}$, β and driving Lévy process M by

$$V_t = \beta + \mathbf{b}'\mathbf{X}_t, \quad t \geq 0,$$

where the state process $\mathbf{X} = (\mathbf{X}_t)_{t\geq 0}$ is the unique solution of the stochastic differential equation

$$\mathrm{d}\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-}\,\mathrm{d}t + \mathbf{e}V_{t-}\,d[M, M]_t^{(d)} = \mathbf{A}\mathbf{X}_{t-}\,\mathrm{d}t + \mathbf{e}(\beta + \mathbf{b}'\mathbf{X}_{t-})\,d[M, M]_t^{(d)},$$

with initial value $\mathbf{X}_0$, independent of $(M_t)_{t\geq 0}$. If the process $(V_t)_{t\geq 0}$ is non-negative almost surely, then $G = (G_t)_{t\geq 0}$, defined by

$$G_0 = 0, \quad \mathrm{d}G_t = \sqrt{V_{t-}}\,\mathrm{d}M_t,$$

is a *COGARCH*(q, p) *process* with parameters $\mathbf{A}$, $\mathbf{b}$, β and driving Lévy process M.

It can be shown that for $p = q = 1$, this definition is equivalent to the definition of the COGARCH(1,1) process given before. The study by Brockwell et al. (2006) gives sufficient conditions for the existence of a strictly stationary solution $(V_t)_{t\geq 0}$ and its positivity and shows that $(V_t)_{t\geq 0}$ has the same autocorrelation structure as a CARMA$(p, q-1)$ process. Hence, the COGARCH(q, p) process allows a more flexible autocorrelation structure than the COGARCH(1,1) process. Under suitable conditions, which among others require M_1 to have expectation zero, it is further shown that nonoverlapping increments of G are uncorrelated, while their squares are not. More precisely,

$$\mathrm{Cov}((G_t - G_{t-r})^2, (G_{t+h} - G_{t+h-r})^2) = \mathbf{b}'\mathrm{e}^{(A+EM_1^2\mathbf{e}\mathbf{b}')h}H_r, \quad h \geq r > 0,$$

where $H_r \in \mathbb{C}^p$ is independent of h. In particular, the squared increments have the covariance structure of a CARMA process.

Acknowledgments

Support of Peter Brockwell's work by National Science Foundation Grant DMS-1107031 and of Alexander Lindner's work by an NTH bottom up project of the state of Lower Saxony is gratefully acknowledged.

References

Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, O., 2001. The distribution of exchange rate volatility. J. Am. Stat. Assoc. 96, 42–55.

Andersen, T.G., Davis, R.A., Kreiss, J.-P., Mikosch, T. (Eds.), 2009. Handbook of Financial Time Series. Springer-Verlag, Berlin.

Andersen, T.G., Benzoni, L., 2009. Realized volatility. In: Andersen, T.G., Davis, R.A., Kreiss, J.-P., Mikosch, T. (Eds.), Handbook of Financial Time Series. Springer-Verlag, Berlin, pp. 555–575.

Applebaum, D., 2004. Lévy Processes and Stochastic Calculus. Cambridge University Press, Cambridge.

Barndorff-Nielsen, O.E., Shephard, N., 2001. Non-Gaussian Ornstein-Uhlenbeck based models and some of their uses in financial economics (with discussion). J. Roy. Stat. Soc., Ser. B 63, 167–241.

Behme, A.D., 2011a. Distributional properties of stationary solutions of $dV_t = V_{t-}dU_t + dL_t$ with Lévy noise. Adv. Appl. Probab. 43, 688–711.

Behme, A.D., 2011b. Generalized Ornstein–Uhlenbeck Processes and Extensions. Ph.D. Thesis, TU Braunschweig.

Behme, A., Lindner, A., 2012. Multivariate generalized Ornstein–Uhlenbeck processes. Stoch. Proc. Appl., to appear, http://dx.doi.org/10.1016/j.spa.2012.01.002.

Behme, A., Lindner, A., Maller, R., 2011. Stationary solutions of the stochastic differential equation $dV_t = V_{t-}dU_t + dL_t$ with Lévy noise. Stoch. Proc. Appl. 121, 91–108.

Bertoin, J., 1996. Lévy Processes. Cambridge University Press, Cambridge.

Bingham, N.H., Goldie, C.M., Teugels, J.L., 1987. Regular Variation. Cambridge University Press, Cambridge.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econom. 31, 307–327.

Brockwell, P.J., 2001. Lévy-driven CARMA processes. Ann. Inst. Stat. Math. 53, 113–124.

Brockwell, P.J., Chadraa, E., Lindner, A., 2006. Continuous-time GARCH processes. Ann. Appl. Probab. 16, 790–826.

Brockwell, P.J., Davis, R.A., 2001. Discussion of "Non-Gaussian Ornstein–Uhlenbeck based models and some of their uses in financial economics," by O.E. Barndorff-Nielsen and N. Shephard, J. Roy. Stat. Soc., Ser. B 63, 218–219.

Brockwell, P.J., Davis, R.A., Yang, Y., 2011. Estimation for non-negative Lévy-driven CARMA processes. J. Bus. Econom. Stat. 29, 250–259.

Brockwell, P.J., Lindner, A., 2009. Existence and uniqueness of stationary Lévy-driven CARMA processes. Stoch. Proc. Appl. 119, 2660–2681.

Brockwell, P.J., Lindner, A., 2012. Integration of CARMA processes and spot volatility modelling. Submitted.

Carmona, P., Petit, F., Yor, M., 1997. On the distribution and asymptotic results for exponential functionals of Lévy processes. In: Yor, M. (Ed.), Exponential Functionals and Prinicipal Values Related to Brownian Motion, Biblioteca de le Revista Mathemàtica Iberoamericana, pp. 73–130.

Cont, R., Tankov, P., 2004. Financial Modelling with Jump Processes. Chapman & Hall/CRC, Boca Raton.

Duan, J.-C., 1997. Augmented GARCH(p, q) process and its diffusion limit. Econometrics 79, 97–127.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50, 987–1008.

Erickson, K.B., Maller, R.A., 2004. Generalised Ornstein–Uhlenbeck processes and the convergence of Lévy integrals. In: Émery, M., Ledoux, M., Yor, M. (Eds.), Séminaire de Probabilités XXXVIII, Lecture Notes in Mathematics 1857, Springer, Berlin, pp. 70–94.

Fasen, V., Klüppelberg, C., Lindner, A., 2006. Extremal behavior of stochastic volatility models. In: Shiryaev, A., Grossinho, M.D.R., Oliviera, P., Esquivel, M. (Eds.), Stochastic Finance, Springer, New York, pp. 107–155.

Goldie, C., 1991. Implicit renewal theory and tails of solutions of random equations. Ann. Appl. Probab. 1, 126–166.

de Haan, L., Karandikar, R.L., 1989. Embedding a stochastic difference equation in a continuous-time process. Stoch. Proc. Appl. 32, 225–235.

Haug, S., Klüppelberg, C., Lindner, A., Zapp, M., 2007. Method of moment estimation in the COGA-RCH(1,1) model. Econom. J. 10, 320–341.

Kesten, H., 1973. Random difference equations and renewal theory for products of random matrices. Acta Math. 131, 207–228.

Klüppelberg, C., Lindner, A., Maller, R., 2004. Stationarity and second order behaviour of discrete and continuous-time GARCH(1,1) processes. J. Appl. Probab. 41, 601–622.

Kyprianou, A.E., 2006. Introductory Lectures on Fluctuations of Lévy Processes with Applications. Springer, Berlin.

Lindner, A., Maller, R., 2005. Lévy integrals and the stationarity of generalised Ornstein-Uhlenbeck processes. Stoch. Proc. Appl. 115, 1701–1722.

Maller, R.A., Müller, G., Szimayer, A., 2008. GARCH modelling in continuous-time for irregularly spaced time series data. Bernoulli 14, 519–542.

Müller, G., 2010. MCMC estimation of the COGARCH(1,1) model. J. Finan. Econom. 8, 481–510.

Nelson, D.B., 1990. ARCH models as diffusion approximations. J. Econom. 45, 7–38.

Paulsen, J., 1993. Risk theory in a stochastic economic environment. Stoch. Proc. Appl. 46, 327–361.

Protter, P.E., 2005. Stochastic Integration and Differential Equations, second ed., Version 2.1, Springer, Berlin.

Sato, K., 1999. Lévy Processes and Infinitely Divisible Distributions. Cambridge University Press, Cambridge.

Schoutens, W., 2003. Lévy Processes in Finance; Pricing Financial Derivatives. John Wiley and Sons Ltd., Chichester.

Stelzer, R., 2010. Multivariate COGARCH(1,1) processes. Bernoulli 16, 80–115.

Todorov, V., 2010. Econometric analysis of jump-driven stochastic volatility models. J. Econom. 160, 12–21.

Todorov, V., Tauchen, G., 2006. Simulation methods for Lévy-driven CARMA stochastic volatility models. J. Bus. Econom. Stat. 24, 455–469.

Tsai, H., Chan, K.S., 2005. A note on non-negative continuous-time processes. J. Roy. Stat. Soc., Ser. B 67, 589–597.

Yoeurp, Ch., Yor, M., 1977. Espace orthogonal à une semimartingale: applications, Unpublished.

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00020-X

Discrete and Continuous Time Extremes of Stationary Processes

K.F. Turkman

Departamento de Estatstica e Investigação Operacional, Universidade de Lisboa, Lisboa 1749-016, Portugal

Abstract

In many applications, the primary interest is the supremum of some continuous-time process over a specified period. However, data are usually available over a discrete set of times and the inference can only be made for the maximum of the process over this discrete set of times. If we want to estimate the extremes of the continuous-time process based on discrete time data, we need to understand the relationship between the continuous and discrete extremes. Thus, we look at asymptotic joint distributions of the maxima of stationary processes and their discrete versions.

Keywords: Extremes, stationary processes, Pickands' grid.

AMS subject Classification: 60G70

1. Introduction

In many applications, the primary interest is the supremum of some continuous-time process over a specified period. However, data are usually available over a discrete set of times and the inference can only be made for the maximum of the process over this discrete set of times. The continuous-time maximum will be larger than the maxima of discrete versions sampled at different frequencies, and if we want to estimate the extremes of the continuous-time process based on discrete-time data, we need to make an adjustment to allow for the effect of discrete sampling and provide a measure of how much smaller it tends to be. In this chapter, we make a review of the most relevant results.

Robinson and Tawn (2000) were the first to point out the importance of sampling spacing on the extremal properties of a process observed at regular discrete-time points. They showed this effect on the extremal indices of the discrete subsequences based on the following arguments:

Let $X(t), t \geq 0$ be a stationary process with distribution $F(x)$. For any time interval $[0, T]$, let $X_\delta(i) = X(i\delta), i = 0, 1, \ldots, T/\delta$, be a subsequence observed at spacing δ. Let

$$M(T) = \sup_{t \in [0,T]} X(t),$$
$$M_\delta(T) = \max_{0 \leq i \leq T/\delta} X(i\delta),$$

be the respective maxima of the continuous process and its subsequence sampled at the δ spacing.

If for any two subsequences sampled at spacings δ and ϵ, the maxima converge with suitable normalization, then for large T,

$$P(M_\delta(T) \leq x) \sim F^{[T/\delta]\theta_\delta}(x) = G_\delta(x),$$
$$P(M_\epsilon(T) \leq x) \sim F^{[T/\epsilon]\theta_\epsilon}(x) = G_\epsilon(x),$$

where $\theta_\delta \in (0, 1]$ and $\theta_\epsilon \in [0, 1]$ are extremal indices of the sequences $X_\delta(i\delta)$ and $X_\epsilon(j\epsilon)$. Since

$$G_\epsilon(x) \sim G_\delta^{(\theta_\epsilon \delta)/(\epsilon\theta_\delta)}(x), \tag{1}$$

one can relate the extremal properties of subsamples at different sampling intervals through the extremal indices. Scotto et al. (2003) obtain more precise limit results from a point process formulation, which exemplify the findings of Robinson and Tawn (2000) and offer more details for a particular class of time series.

In order to relate the asymptotic distribution of $M(T)$ to $M_\delta(T)$, for some fixed sampling spacing δ, it is tempting to use the above arguments and argue that as $\epsilon \to 0$, if

$$\lim_{\epsilon \to 0} \theta_\epsilon / \epsilon = H,$$

then (taking $\delta = 1$ and $\theta_\delta = \theta$)

$$P(M(T) \leq x) \sim G_1^{H/\theta}(x).$$

In this case, H/θ may be seen as the adjustment needed to allow for using a discrete subsequence.

Based on the assumption that for most continuous processes, there is a fixed sampling spacing ϵ, for which $M_\epsilon(T)$ is a sufficiently good approximation to $M(T)$ over the time interval $[0, T]$, Anderson (2003) suggests using

$$\phi = \frac{\theta_\epsilon \delta}{\epsilon \theta_\delta}, \tag{2}$$

as the adjustment for using discrete sampling. He further shows that

$$1 \leq \phi \leq \frac{1}{P(S \geq \delta)}, \tag{3}$$

where S is the excursion time of the continuous process above a high level, which he calls the storm period. The probability that appears as the upper bound in (3) cannot be estimated from the discrete δ-observations, but if one can give a conservative estimate for the probability of storm duration being longer than the sampling spacing δ, for example, by eliciting expert opinion, then as Anderson (2003) suggests, the inequality (3) can help to relate important quantities such as the return levels calculated at different sampling spacings. If $x_{n,\epsilon}$, $x_{n,\delta}$ are, respectively, n time units return levels based on ϵ and δ-observations, then

$$n(1 - G_\delta(x_{n,\delta})) = 1,$$
$$n(1 - G_\epsilon(x_{n,\epsilon})) = n(1 - G_\delta^\phi(x_{n,\epsilon})) = 1,$$

and $x_{n,\epsilon} = x_{n\phi,\delta}$. Thus, if ϵ sampling is sufficiently dense so that $M_\epsilon(h)$ is a sufficiently good approximation to $M(h)$, then $x_{n\phi,\delta}$ can be taken as the n time units return level from the continuous observations.

The above approximations and bounds suggested by Anderson (2003) depend strongly on the assumption that one can approximate a continuous maximum in terms of a discrete maximum over a fixed sampling spacing to a desired level of accuracy. Therefore, one needs to understand the relation between the maxima of continuous- and discrete-time processes in order to judge the robustness of the arguments given by Anderson (2003). Further, such results may help to find sharper bounds for the adjustment.

The coarsest grid over which continuous and discrete maxima over fixed intervals have the same asymptotic distribution is fundamental in obtaining limit results. Such grid is defined as a family of uniformly spaced grids with grid spacing converging to zero at a specified rate and is called the Pickands' grid. A standard Pickands' grid relating continuous and discrete maxima over fixed intervals is suggested by Leadbetter et al. (1983), see also Albin (1987, 1990) and Piterbarg (2004). Although, this grid adapted to $X(t)$ requires hard to verify technical conditions, it permits elegant characterization of results. This family of grids (we refer to them by their spacing δ) that will be defined formally in Section 2, is taken as a function of the high threshold u such that

$$\delta_a = \delta(a, u) = \{aq(u)j, j = 0, 1, 2, \dots\},$$

where $a > 0$ is an arbitrarily small positive real number and $q(u)$ is a sequence typically converging to 0 as $u \to \infty$. Here, a is a constant regulating the rate of convergence of the grid spacing to 0. As Leadbetter et al. (1983) suggest, one can define a universal grid δ without the extra parametrization depending on a, but this extra parametrization brings flexibility in proofs. Note that, as the threshold u tends to infinity, the excursion times of the continuous process above this high level get shorter and in order to capture these events by a discrete version of the process, corresponding grid spacing needs to converge to zero accordingly. $q(u)$ quantifies this relationship.

We will call any other grid $\delta_b = \{jbg(u), j = 0, 1, 2, \ldots\}$ a sparse or loose grid if $u \to \infty$, $q(u) = o(g(u))$, Pickands' grid if $q(u) = o(g(u))$, and dense grid if $g(u) = o(q(u))$. If δ_a is a Pickands' grid, then the limit as $a \to 0$, δ_a is a dense grid.

Conditions for the existence of the limit over increasing intervals

$$\lim_{T\to\infty} P(M(T) \le u_T(x)) = G(x),$$

with suitable linear normalization $u_T(x) = a_T + b_T x$ and its relation to the distribution of maximum over a Pickands' grid are well known. In this case, the high level has to be chosen as a function of the increasing interval and consequently, the Pickands' grid is also a function of the increasing time interval; see Leadbetter et al. (1983) for Gaussian processes, and Albin (1987, 1990) for general stationary processes. However, the most complete characterization of the relation between discrete- and continuous-time extremes is given by Piterbarg (2004) for stationary Gaussian processes; see Husler (2004) for the generalization to locally stationary Gaussian processes. Typically, for a grid δ_a with suitable normalizations, the asymptotic joint distribution

$$P(M(T) \le u_T(x), M_{\delta_a}(T) \le u_{T,\delta_a}(y)) \tag{4}$$

as $T \to \infty$, is studied for three distinct cases:

- When the sampling is done over a dense grid, then (4) is studied for a Pickands' grid δ_a, and as $a \to 0$, the bivariate limiting distribution given in (4) is degenerate, converging to one of the identical marginal distributions of the coinciding continuous- and discrete-time maxima. Thus

$$\lim_{a\to 0}\lim_{T\to\infty} P(M(T) \le u_T(x), M_{\delta_a}(T) \le u_{T,\delta_a}(y)) = G(z),$$

  where $z = \min(x, y)$.
- When the sampling is done over a Pickands' grid, then the maxima are asymptotically dependent, and the limit of (4) for fixed a is given by

$$G(x)G_{\delta_a}(y)G(a, x, y), \tag{5}$$

  where $G(x)$ and $G_{\delta_a}(y)$ are the respective limiting marginal distributions of the continuous- and discrete-time maxima and $G(a, x, y)$ is a function explaining the degree of asymptotic dependence of the respective maxima.
- If the sampling is done over a sparse grid, then typically the maxima of the continuous- and discrete-maxima grow with different rates, but it is still possible to find suitable sets of normalization $u_{T,\delta_a}(x)$ and $u_T(y)$, such that the normalized maxima are asymptotically independent, yet having nondegenerate asymptotic marginal distributions.

The function $G(a, x, y)$ in (5) is calculated by Piterbarg (2004) for Gaussian processes. Specifically, if $X(t)$ is a zero mean stationary Gaussian process with unit variance and covariance function $r(t)$ satisfying for some $\alpha > 0$,

$$r(t) = 1 - |t|^\alpha + o(|t|^\alpha),$$

as $t \to 0$ with $r(t) < 1$ for all $t > 0$, and further as $t \to \infty$, $r(t) = o(1/\log t)$ then, with suitable normalization

$$\begin{aligned} P(M(T) \le u_T(x), M_{\delta_a}(T) \le u_{T,\delta_a}(y)) \\ &= G(x)G_{\delta_a}(y)G(a,x,y) \\ &= \exp(-e^{-x})\exp(-e^{-y})\exp(-G_a(\log H + x, \log H_a + y)). \end{aligned} \tag{6}$$

Here, $0 < G_a(x,y) < \infty$ appears as the limit

$$G_a(x,y) = \lim_{T\to\infty} \frac{1}{T} G_a(x,y,T),$$

with

$$G_a(x,y,T) = \int_{-\infty}^{\infty} e^{\upsilon} P\left(\max_{k:ka\in[0,T]} \sqrt{2}B_{\alpha/2}(ka) > \upsilon, \max_{k:t\in[0,T]} \sqrt{2}B_{\alpha/2}(t) - t^{\alpha} > \upsilon + y \right) d\upsilon. \tag{7}$$

The process $B_{\alpha/2}(t)$ that appears in the expression (7) is the fractional Brownian motion with variance $|t|^{\alpha}$, whereas H, H_a are Pickands' constants that appear in the marginal limiting distributions, see, for example, Leadbetter et al. (1983) or Piterbarg (2004).

In the Gaussian case, all asymptotic results on extremes can conveniently be characterized by the covariance function and the proofs are constructed around this tool, see, for example, Husler (1999). For non-Gaussian processes, different sets of conditions are needed. In Section 2, we report similar results for the joint asymptotic distribution of continuous–discrete maxima of stationary, but not necessarily Gaussian processes. These results are constructed on the assumptions and techniques of Albin (1987, 1990). We will first look at the asymptotic joint distribution of maxima sampled over a Pickands' grid δ_a and any other grid δ_b, namely

$$P(M_{\delta_a}(h) \le u, M_{\delta_b}(h) \le u'),$$

in a fixed interval $[0,h]$, for some suitably chosen and increasing levels u and u', then extend the results to increasing time intervals. For relative ease in notation, we report the results for stationary processes with regularly varying tails, but following Albin (1990), it is possible to extend the results for other types of tail behavior.

Clearly, the rate at which the Pickands' grid tends to 0 will depend on sample path properties of the continuous process as well as the tail behavior of its marginal distribution. Here we give some examples:

1. If $X(t)$ is a stationary, 0 mean Gaussian process with covariance function

$$r(t) = 1 - C|t|^{\alpha} + o(|t|^{\alpha}),$$

for some $\alpha \in [0,2]$, and $C > 0$, as $t \to 0$, then the Pickands' grid is chosen as

$$\delta_a = \{jaq(u), j = 0, 1, 2, ..\}$$

with

$$q(u) = u^{-2/\alpha},$$

so that

$$\lim_{a \downarrow 0} \lim_{u \to \infty} |P(M(h) > u) - P(M_{\delta_a}(h) > u)| = 0, \tag{8}$$

see Piterbarg (2004) and Berman (1982).

2. If $X(t)$ is a standardized differentiable stationary Gaussian process satisfying

$$r(t) = 1 + \frac{1}{2} r''(0) t^2 + o(t^2),$$

as $t \to 0$ and $Y(t)$ is the moving $\mathcal{L}^2$-norm process given by

$$Y(t) = \int_t^{1+t} X^2(s) ds,$$

then the grid ϵ can be chosen with

$$q(u) = (1 \vee u)^{-1/2}$$

(Albin, 2001).

3. If $X(t)$ is an α-stable process with $\alpha > 1$, then it is possible to take $q(u) = 1$, and (8) will hold with $a \to 0$ (Hsing and Leadbetter, 1998; Samorodnitsky and Taqqu, 1994).
4. On the other hand, If $X(t)$ is a moving average of an α-stable process, with $\alpha > 1$, then (8) will hold for a fixed grid ϵ, that is, it will hold with $q(u) = 1$ and for any $a > 0$ (Albin, 2001).
5. If $X(t)$ is an α-stable process with $\alpha < 1$ then (8) holds with

$$q(u) = (-u)^{\alpha/[2(1-\alpha)]},$$

as $a \to 0$ (Albin, 2001).

6. If $X_i(t)$ are independent standardized stationary processes with covariance functions satisfying

$$1 + \frac{1}{2} C_i |t|^\alpha + o(|t|^\alpha),$$

as $t \to 0$, for some $C_i > 0, \alpha \in [0,2]$,

$$Z(t) = \sum_{i=1}^{m} X_i^2(t),$$

then (8) will hold with

$$q(u) = u^{-1/\alpha},$$

as $a \to 0$ (Albin, 1987).

It is clear that except for some special processes such as the moving average α-stable process with $\alpha > 1$, it may not be possible to find a fixed sampling spacing, for which the discrete maximum approximates the continuous maximum to a desired level. Note that the continuous maximum is almost surely larger than the discrete maximum and the function $q(u)$ also quantifies the relative size of each maxima through the relation

$$q(u) \sim \frac{P(X(0) > u)}{P(M(0,1) > u)},$$

as $u \to \infty$, see Hsing and Leadbetter (1998).

In the next section, we state the technical conditions as well as the main results for the marginal convergence of maxima of stationary process with heavy tailed distributions, which will help in understanding the asymptotic convergence of the joint distributions of continuous- and discrete-time extremes. The proofs will be omitted, as they can be found in the works of Albin (1987, 1990). The conditions and the results will be grouped under subsections, first for results on finite intervals, then on increasing intervals. The proofs of the new results on the asymptotic joint distributions of continuous- and discrete-time extremes given in Theorems 2 and 5 are tedious, and therefore will not be given here. The detailed arguments can be found in the study by Turkman (2011). In Section 3, we give some asymptotic results on the maxima of the periodogram of a Gaussian time series, calculated over Fourier frequencies $\omega_j = 2\pi j/n$, $j = 1, 2, \ldots, [\frac{1}{2}(n-1)]$ and continuous frequencies in $[0, \pi]$ to highlight these technical results.

2. Conditions and main results

2.1. Finite intervals

Assume that the stationary process $X(t)$ satisfies the following sufficient conditions of Albin (1987, 1990) for the marginal convergence of the continuous maximum over finite intervals:

1. *Condition C1*
   F belongs to the Frechét domain of attraction so that for any $x > 0$,

$$\lim_{u \to \infty} \frac{1 - F(ux)}{1 - F(u)} = x^{-c},$$

   for some $c > 0$.
2. *Condition C2*

For a strictly positive function $q = q(u)$, let $\delta_a = \{aq(u)j, j = 0, 1, 2, \ldots, [h/aq]\}$ be a grid over the interval $[0, h]$ such that (writing for simplicity $q = q(u)$)

$$\limsup_{u\to\infty} \frac{q(u)}{1 - F(u)} P\left(M(h) > u, \max_{a\le aqj\le h} X(aqj) \le u\right) = 0, \tag{9}$$

as $a \to 0$. For any fixed, but sufficiently small $a > 0$, we call δ_a that makes the discrete approximations sufficiently accurate in the sense given in (9) as *Pickands' grid*. On the other hand, any Pickands' grid with $a \to 0$ will be called dense grid.

3. *Condition C3*
   Assume that there exist a sequence of random variables $\{\eta_{a,x}(k)\}_{k=1}^{\infty}$, and a strictly positive function $q(u)$ with $\lim_{u\to\infty} q(u) = 0$ such that for all $x \ge 1$ and for all $a > 0$, and for any finite integer N, as $u \to \infty$,

$$\left(\frac{1}{u}X(aq), \ldots, \frac{1}{u}X(aqN) | \frac{1}{u}X(0) > x\right) \to^{D} (\eta_{a,x}(1), \ldots, \eta_{a,x}(N)). \tag{10}$$

4. Condition *C4, Short-lasting exceedances*

$$\limsup_{u\to\infty} \frac{1}{1 - F(u)} \sum_{k=N}^{[h/aq]} P(X(0) > u, X(aqk) > u) \to 0,$$

as $N \to \infty$, For all fixed $a > 0$.

We refer the reader to the works of Albin (1990) for the details of these assumptions. Condition *C3* is a natural extension of the condition *C1* and it is satisfied by most processes. Albin (1990) gives an alternative condition to *C2*, which can be verified by two-dimensional distributions of the process. However, we note that condition *C4* is not always satisfied. We refer to Albin (1987) and Husler et al. (2010) for asymptotic results when this condition is violated.

THEOREM 1. (Marginal convergence of maxima over Pickands' or denser grids by Albin (1987))

1. *Assume that conditions C1, C3, and C4 are satisfied. Then for any $a > 0$ fixed,*

$$\lim_{u\to\infty} \frac{q(u)}{1 - F(u)} P(M_\delta(h) > u) = hH_{a,1}(1),$$

and for any $x > 0$

$$\lim_{u\to\infty} \frac{q(u)}{1 - F(ux)} P(M_\delta(h) > ux) = hH_{a,x}(x)$$

where

$$H_{a,x}(x) = \frac{1}{a} P\left(\max_{k \geq 1} \eta_{a,x} \leq x \right), \tag{11}$$

and

$$\lim_{a \to 0} H_{a,x}(x) = H_x(x), \tag{12}$$

exist with $0 < H_x(x) < \infty$.

2. *If further, condition C2 is satisfied, then*

$$\lim_{u \to \infty} \frac{q(u)}{1 - F(u)} P(M(h) > u) = h H_1(1), \tag{13}$$

and

$$\lim_{u \to \infty} \frac{q(u)}{1 - F(ux))} P(M(h) > ux) = h H_x(x),$$

so that

$$\lim_{u \to \infty} \frac{q(u)}{1 - F(u)} P(M(h) > ux) = h H_x(x) x^{-c}. \tag{14}$$

Note that $H_x(x)$ *is not a constant, and therefore (14) may indicate that the distribution functions of* $M(h)$ *and* F *may not belong to the same domain of attraction. However, Albin (1990) shows that*

$$\frac{q(ux)}{q(u)} = x^{-c^*}, \tag{15}$$

for some $c^* \in [0, c)$, *for all* $x > 0$ *so that as* $u \to \infty$,

$$\lim_{u \to \infty} \frac{q(u)}{1 - F(u)} P(M(h) > ux) = h H_1(1) x^{-(c - c^*)},$$

for some $c^* \in [0, c)$. *Hence, the distribution functions of* $M(h)$ *and* F *belong to Frechét domain of attraction, having different shape parameters.*

Condition C3 is given in terms of conditioning on the event $\{X(0) > ux\}$. *However, an alternative formulation in terms of conditioning on the event* $\{X(0) = ux\}$ *can also be given:*

COROLLARY 1. *Assume further that* F *has a density* f *satisfying*

$$\lim_{u \to \infty} \frac{u f(u)}{1 - F(u)} = c,$$

for some $c > 0$ and assume further that there exists variables $\{\zeta_{a,x}(k)\}_{k=1}^{\infty}$ such that

$$\left(\frac{1}{u}X(aq),\ldots,\frac{1}{u}X(Naq)|X(0)=ux\right)\to^{D}\{\zeta_{a,x}(k)\}_{k=1}^{N}, \tag{16}$$

for all $X > 1$ and for all N. Then (13) and (14) hold with

$$H_x(x)=\lim_{a\to 0}\frac{1}{a}\int_1^{\infty}P\left(\max_{k\geq 1}\zeta_{a,xy}(k)\leq x\right)cy^{-(c+1)}dy,$$

and

$$H_1(1)=\lim_{a\to 0}\frac{1}{a}\int_1^{\infty}P\left(\max_{k\geq 1}\zeta_{a,y}(k)\leq 1\right)cy^{-(c+1)}dy.$$

Equipped with the results for marginal convergence, we can now state the results for joint convergence:

THEOREM 2. (Joint convergence of maxima over Pickands' or denser grids)

1. *For any $a > 0$, $b > 0$, such that $a < b$, Let*

$$\delta_a=\{jaq(u), j=0,1,2,\ldots,[h/aq]\}$$

and

$$\delta_b=\{jbq(u), j=0,1,2,\ldots,[h/bq]\}$$

be two Pickands' grids satisfying conditions C1–C3. Let $z=\min(x,y)(=x\wedge y)$ and $v=\max(x,y)(=x\vee y)$.
Then

$$\lim_{u\to\infty}\frac{q(u)}{1-F(u)}P(\{M_{\delta_a}(h)>ux\}\cup\{M_{\delta_b}(h)>uy\})=hH_{a,b,z}(x,y)z^{-c},$$

where

$$H_{a,b,z}(x,y)=\frac{1}{a}P\left(\max_{k\geq 1}\eta_{a,z}(k)\leq x,\max_{k\geq 1}\eta_{b,z}(k)\leq y\right).$$

2.

$$\lim_{a\to 0}H_z(a,b,x,y)=H_z(b,x,y),$$

exists with $0 < H_z(b,x,y) < \infty$ and

$$\lim_{u\to\infty}\frac{q(u)}{1-F(u)}P(M(h)>ux\cup M_{\delta_b}(h)>uy)=hH_z(b,x,y)z^{-c},$$

where

$$H_z(b,x,y) = \lim_{a\to 0} \frac{1}{a} P\left(\max_{i\geq 1} \eta_{a,z}(i) \leq x, \max_{j\geq 1} \eta_{b,z}(j) \leq y\right).$$

3. *The limit*

$$\lim_{b\to 0} H_z(b,x,y) = H_z(z)$$

exists with $0 < H_z(z) < \infty$, *where*

$$H_z(z) = \lim_{b\to 0} \frac{1}{b} P\left(\max_{i\geq 1} \eta_{z,b}(i) \leq z\right),$$

and hence

$$\lim_{b\to 0} \lim_{u\to\infty} \frac{q(u)}{1-F(u)} P(M(h) > ux \cup M_{\delta_b}(h) > uy) = hH_z(z)z^{-c}.$$

The proof is quite tedious and is based on finding asymptotic bounds for the expression

$$P(\{M_{\delta_a}(I_0) > uy\} \cup \{M_{\delta_b}(I_0) > ux\}),$$

where $I_0 = [0, aq, 2aq, \ldots, Naq)$ for some integer N. This result extends the proof of Theorem 1 of Albin (1990), where asymptotic bounds for the expression

$$P(\{M_{\delta_a}(I_0) > uy\}$$

are derived, see Turkman (2011) for details.

We now look at the asymptotic independence of maxima calculated over Pickands' and sparse grids.

For some strictly positive function $g = g(u)$ such that $\lim_{u\to\infty} g(u) = 0$ and

$$\lim_{u\to\infty} \frac{q(u)}{g(u)} = 0,$$

let

$$\delta_b = \{kbg(u), k = 0, 1, 2, \ldots, [h/bg]\}, \tag{17}$$

be a sparse grid (with respect to the Pickands' grid). Let

$$u' = \left(\frac{q(u)}{g(u)}\right)^{1/c} u, \tag{18}$$

so that as $u \to \infty$, $u' = o(u) \to \infty$. Further assume that $g(u)$ is such that the slowly varying function L in the representation $1 - F(x) = x^{-c}L(x)$ satisfies the condition

$$\lim_{u\to\infty} \frac{L(u')}{L(u)} = 1.$$

Assume that there exists variables $\{\zeta_{b,y}(k)\}_{k=1}^{\infty}$ such that for any $y > 0$ and for any N,

$$\left(\frac{1}{u'}X(bg), \ldots, \frac{1}{u'}X(Nbg) | \frac{1}{u'}X(0) > y\right) \to^{D} (\zeta_{b,y}(1), \ldots, \zeta_{b,y}(N)).$$

THEOREM 3. (Joint convergence of maxima over Pickands' and sparse grids)

For any Pickands' grid δ_a and sparse grid δ_b defined in (17) and for any $x > 0$, $y > 0$,

1.

$$\lim_{u\to\infty} \frac{q(u)}{1 - F(u)} P(M_{\delta_a}(h) \geq uy, M_{\delta_b}(h) \geq u'x) = 0.$$

2.

$$\lim_{u\to\infty} \frac{q(u)}{1 - F(u)} P(M(h) > uy \cup M_{\delta_b}(h) > u'x) = hy^{-c}H_y(y) + hx^{-c}H'_x(x),$$

where, $0 < H'_x(x) < \infty$ is the limit

$$H'_x(x) = \lim_{b\to 0} \frac{1}{b} P\left(\max_{k\geq 1} \zeta_{b,x}(k) \leq x\right),$$

and $H_y(y)$ is given in (12).

2.2. *Increasing intervals*

Let

$$M(T) = \max_{t\in[0,T]} X(t),$$
$$M_\delta(T) = \max_{0\leq jaq\leq T} X(jaq),$$

and u_T be chosen such that as $T \to \infty$,

$$\frac{T}{q(u_T)}(1 - F(u_T)) = 1.$$

For simplicity in notation, let $q = q(u_T)$.

Assume that

1. *Condition* $\Delta(u_{T,1}(x_1), u_{T,2}(x_2))$
   For $0 < s < t < T$ and x_i, $i = 1, 2$ write

$$\Im^T_{s,t}(x_1, x_2) = \sigma\{(X(v) \leq u_{T,j}(x_i) : x_i > 0, s \leq v \leq t, i = 1, 2, j = 1, 2\},$$

the sigma field generated by the respective events and

$$\alpha_{T,l}(x_1, x_2) = \sup\{|P(AB) - P(A)P(B)| : A \in \Im_{0,s}^{T}(x_1, x_2), B \in \Im_{s+l,t}^{T}(x_1, x_2), s \geq 0, l + s \leq T\}.$$

$\Delta(u_{T,1}(x_1), u_{T,2}(x_2)$ is said to hold for the process $X(t)$ and the family of pair of constants $\{u_{T,1}(x_1), u_{T,2}(x_2)\}$, if $\alpha_{T,l}(x_1, x_2) \to 0$, as $T \to \infty$ for some $l_T = o(T)$. Note that this is a variation of the usual $D(u_n)$ condition, adapted to events generated by two different normalization, see Mladenovic and Piterbarg (2006) for a similar condition.

2. Assume that the X-process satisfies the *No clusters of clusters condition* of Albin (1990): This condition is said to hold for $X(t)$ with respect to the grid $\delta = \{jaq(u), j = 0, 1, 2, \ldots .\}$ if for any finite $h > 0$

$$\limsup_{u\to\infty} \frac{1}{1 - F(u)} \sum_{\frac{1}{2}h < jaq < \epsilon T} P(X(0) > u, X(jaq) > u) \to 0, \quad (19)$$

as $\epsilon \to 0$.

THEOREM 4. Maxima over increasing intervals

Assume that the X-process satisfies the conditions of the previous section as well as the conditions $\Delta(u_T x, u_T y)$ and (19). Then,

1. *For any Pickands' grid δ_b given in Theorem 2,*

$$\lim_{T\to\infty} P(M(T) \leq u_T x, M_{\delta_b}(T) \leq u_T y) = \exp[-z^{-c} H_z(b, x, y)].$$

2. *For any sparse grid δ_b defined as in Theorem 4, and*

$$u'_T = \left(\frac{q(u_T)}{g(u_T)}\right)^{1/c} u_T,$$

assume that the process satisfies the $\Delta(u_T x, u'_T y)$ condition as well as the no clusters of clusters condition given by (19). Then

$$\lim_{T\to\infty} P(M(T) \leq u_T x, M_{\delta_b}(T) \leq u'_T y) = \exp[-x^{-c} H_x(x) - y^{-c} H'_y(y)]. \quad (20)$$

It is possible to extend Corollary 1 to joint convergence: If δ_b is a Pickands' grid given in (1) of Theorem 2, then under the alternative conditioning (16).

COROLLARY 2.

$$\lim_{T\to\infty} P(M(T) \leq ux, M_{\delta_b}(T) \leq uy) = \exp[-z^{-c} \hat{H}_z(b, x, y)],$$

where,

$$\hat{H}_z(x,y) = \lim_{a\to 0} \frac{1}{a} \int_1^\infty P\left(\max_{k\geq 1} \zeta_{a,zw}(k) \leq x, \max_{k\geq 1} \zeta_{b,zw}(k) \leq y\right) c w^{-(c+1)} dw.$$

Further,

$$\lim_{b\to 0} \hat{H}_z(b,x,y) = \hat{H}_z(z). \tag{21}$$

For ease of notation, we have given the results for distributions in the domain of attraction of Frechét. However, with some standard changes, it is possible to extend the results to other domains of attraction, see, for example, Albin (1987, 1990) for conditions and proofs of marginal convergence.

It is difficult to verify the conditions and the specific expressions given for $H_x(x)$ and $H_z(b,x,y)$ for processes other than Gaussian processes. However, there are some processes that are transformations of Gaussian processes such as the Rayleigh process for which these conditions may be verified and the specific expressions may be calculated, see Albin (1990) for details. Here, we give another example for which it is possible to obtain specific results.

3. Periodogram

Let $\{X_t\}_{t=1}^n$ be a stationary time series with 0 mean and finite variance. The periodogram

$$\begin{aligned} I_n(\omega) &= \frac{2}{n}\left|\sum_{t=1}^n X_t e^{i\omega t}\right|^2 \\ &= X_n^2(\omega) + Y_n^2(\omega), \quad \omega \in [0,\pi] \end{aligned} \tag{22}$$

where

$$X_n(\omega) = \sqrt{2/n} \sum_{t=1}^n X_t \cos(\omega t), \tag{23}$$

$$Y_n(\omega) = \sqrt{2/n} \sum_{t=1}^n Y_t \sin(\omega t), \tag{24}$$

appears to be the natural estimator of the spectral density function $h(\omega)$, yet it is inconsistent and its erratic behavior is well known. One reason for this erratic behavior is that the maximum of the periodogram over any finite interval diverges as $n \to \infty$ almost surely in the order of $2\log n$ (see, e.g., Turkman and Walker (1990)). Fundamental reason for this erratic behavior is that the correlation functions $r_{n,X}(t)$ and $r_{n,Y}(t)$ of the processes $X_n(\omega)$ and $Y_n(\omega)$ having the behavior

$$1 - \frac{n^2}{3} t + o(t^2),$$

as $t \to 0$. Thus, these processes have second spectral moments that diverge, as $n \to \infty$.

The periodogram plays an important role in tests of hypotheses regarding the jumps in the spectral distribution function. In particular, the maximum of periodogram ordinates over the Fourier frequencies $\omega_j = 2\pi j/n$, $j = 1, 2, \ldots, [\frac{1}{2}(n-1)]$ given by

$$M_{n,I} = \max_{1 \le j \le [\frac{1}{2}(n-1)]} I_n(\omega_j),$$

plays a central role in these tests of hypotheses. The convenience of using this test statistic is that when X_t is a zero mean Gaussian process, the periodogram ordinates over these Fourier frequencies constitute an i.i.d. standard exponential sample, and the asymptotic distribution of the test statistics is relatively easy to obtain. On the other hand, when X_t is a zero mean, finite variance non-Gaussian process, these ordinates are neither independent nor uncorrelated, but Davis and Mikosch (1999) show that the asymptotic distribution of $M_{n,I}$ still has a similar behavior.

In principle, tests on the jumps should be constructed based on the maxima of the periodogram over the continuous range of frequencies

$$M_I = \max_{\omega \in [0,\pi]} I_n(\omega),$$

and it is not very clear how much power, if any, one loses by using the discrete maxima instead of the continuous maxima while constructing these tests. Walker (1965) remarks that indeed greater power can be achieved by using the continuous maxima, indicating that the discrete maximum over the Fourier frequencies may not sufficiently approximate the continuous maximum. In fact, when X_t is a Gaussian sequence, a Pickands' grid for the periodogram can be found using a theorem of Bernstein on trigonometric polynomials (see Turkman and Walker (1984) or Zygmund (1959)):

THEOREM 5. If a trigonometric polynomial of order n,

$$T(x) = \sum_{k=-n}^{n} c_k e^{ikx}$$

satisfies $|T(x)| \le M$ *for every* x *and for some constant* M, *then the derivative* $T'(x)$ *satisfies* $|T'(x)| \le nM$.

Clearly, the periodogram is a trigonometric polynomial of order n. If $\delta_a = \{w_j = jaq(n), j = 0, 1, \ldots\}$ is a partition of $[0, \pi]$, then for any $\omega \in [\omega_j, \omega_{j+1})$, there exists another $\omega^* \in [\omega_j, \omega_{j+1})$ such that

$$I_n(\omega) = I_n(\omega_j) + (\omega - \omega_j) I_n'(w^*).$$

Denote by $M_{\delta_a,I}$, the maximum of the periodogram over the grid δ_a. Then, from Bernstein's theorem, almost surely

$$\begin{aligned} M_{\delta_a,I} \le M_I &\le M_{\delta_a,I} + \max_{\omega_1} |\omega_1 - \omega_j| \max_{\omega \in [0,\pi]} I_n'(\omega) \\ &\le M_{\delta_a,I} + \pi aq(n) n M_I. \end{aligned}$$

Almost surely

$$\lim_{n\to\infty} \frac{M_I}{2\log n} = 1,$$

(see Turkman and Walker (1990)), hence as $n \to \infty$,

$$0 \le M_I - M_{\delta_a,I} \le 2\pi a q(u) n \log n.$$

Therefore, choosing $q(u) = (n \log n)^{-1}$, we see that almost surely

$$\lim_{a\to 0} \lim_{n\to\infty} M_{n,I} - M_{n,\delta_a,I} = 0.$$

These arguments need to be slightly more precise near $\omega = 0$, but we omit the details that can be found in the works of Turkman and Walker (1984).

Thus, the Fourier frequencies $\omega_j = 2\pi j/n$, $j = 1, 2, \ldots, [\frac{1}{2}(n-1)]$ form a sparse grid and the maximum over the Fourier frequencies and over a dense grid grow with different rates. In fact, as $n \to \infty$,

$$P(M_{n,I} \le 2x + \log n - 2\log 2) \to \exp(e^{-x}), \tag{25}$$

whereas,

$$P(M_I \le 2x + 2\log n + \log\log n - \log 3/\pi) \to \exp(e^{-x}). \tag{26}$$

Thus, if Λ is a random variable with standard Gumbel distribution, then

$$M_{n,I} =^d 2\Lambda + 2\log n - 2\log 2,$$

whereas

$$M_I =^d 2\Lambda + 2\log n + \log\log n - \log\frac{3}{\pi},$$

These results show the degree of deviance of the maximum of the periodogram ordinates at Fourier frequencies from the maximum over the continuous range of frequencies, differing in the limit by an order of $\log\log n$.

The limit in (25) is given by Walker (1965), whereas the limit in (26) is obtained by Turkman and Walker (1984) by showing that as $n \to \infty$,

$$P(M_{n,I} > u_n) \sim \mu(u_n),$$

where $\mu(u_n)$ is the upcrossing intensity of the high-level u_n. However, in order to obtain such a result, it is needed to verify that the second-order moment of upcrossings given by

$$E[N_{u_n}(N_{u_n} - 1)]$$

is negligibly small, that is

$$E[N_{u_n}(N_{u_n} - 1)] = o(\mu(u_n)), \tag{27}$$

where N_{u_n} is the number of upcrossings of the level u_n in the interval $[0, \pi]$. Methods that are employed to obtain such results are specific for Gaussian processes and other processes that are simple transformations of Gaussian processes, such as the periodogram for Gaussian time series. It would be interesting to characterize the joint limiting distribution of the periodogram maximum over the continuous range of frequencies and the periodogram maximum over a Pickands' grid in terms of the random sequence $\{\zeta_{a,x}(k)\}_{k=1}^{\infty}$ given in (10). Albin (1990) shows that under the condition (27),

$$\lim_{a\to 0} \frac{1}{a} P(\zeta_{a,x}(1) \le x) = \lim_{a\to 0} \frac{1}{a} P\left(\sup_{k\ge 1} \zeta_{a,x} \le x\right),$$

therefore, the sequence $\{\zeta_{a,x}(k)\}_{k=1}^{\infty}$ must be degenerate in some sense, facilitating part of the tedious technical work.

Characterizations of Section 2 give a very detailed and accurate description of the asymptotic relationship between the discrete and continuous maxima, but they have very limited practical use since the conditions are generally hard to verify and little is known on possible estimators for expressions such as $H_z(b, x, y)$ defining the degree of dependence between the continuous and discrete extremes. Therefore, it is very important to get simpler and more robust representations for the relationship between continuous and discrete maxima, such as the adjustment (3) suggested by Anderson (2003), which are more adapted for statistical inference permitting numerical computations and applications. For statistical applications, most interesting case is the joint distribution of maxima over the continuous range and a sparse grid. However, asymptotic results are not particularly useful, as these maxima are asymptotically independent. Therefore, more refined class of models describing the tails of asymptotically independent distributions at subasymptotic levels are needed. The study of rates of convergence related to the reported asymptotic results may also be very useful in getting sharper bounds for the adjustment given in Eq. (3).

Acknowledgment

This work is partially supported by FCT projects PTDC/MAT/118335/2010 and Pest-OE/MAT/UI0006/2011.

References

Albin, P., 1987. On extremal theory for non differentiable stationary processes. Ph.D. Thesis, Department of Mathematical Statistics, University of Lund.

Albin, P., 1990. On extremal theory for stationary processes. Ann. Probab. 18, 92–128.

Albin, P., 2001. On extremes and streams of upcrossings. Stochastic Process. Appl. 94, 271–300.

Anderson, C.W., 2003. A note on continuous-time extremes from discrete time observations. Unpublished research report.

Berman, S., 1982. Sojourns and extremes of stationary processes. Ann. Probab. 10, 1–46.
Davis, R.A., Mikosch, T., 1999. The maximum of the periodogram of a non-Gaussian sequence. Ann. Probab. 27, 522–536.
Hsing, T., Leadbetter, M.R., 1998. On the excursion random measure of stationary processes. Ann. Probab. 26, 710–742.
Husler, J., 1999. Extremes of a Gaussian process and the constant H_α. Extremes 2, 59–70.
Husler, J., 2004. Dependence between extreme values of discrete and continuous time locally stationary Gaussian processes. Extremes 7, 179–190.
Husler, J., Ladneva, A., Pitebarg, V., 2010. On clusters of high extremes of Gaussian stationary processes with ϵ- separation. Elect. J. of Prob. 15(59): 1825–1862.
Leadbetter, M.R., Rootzén, H., Lindgren, G., 1983. Extreme value theory for continuous parameter stationary processes. Z. Wahrsch. verw. Gebiete 60, 1–20.
Mladenovic, P., Piterbarg, V., 2006. On asymptotic distribution of maxima of complete and incomplete samples from stationary sequences. Stochastic Process. Appl. 116, 1977–1991.
Piterbarg, V., 2004. Discrete and continuous time extremes of Gaussian processes. Extremes 7, 161–177.
Robinson, M.E., Tawn, J., 2000. Extremal analysis of processes sampled at different frequencies. J. R. Stat. Soc. B 62, 117–135.
Samorodnitsky, G., Taqqu, M., 1994. Stable Non-Gaussian Random Processes. Chapman and Hall, Boca Raton.
Scotto, M.G., Turkman, K.F., Anderson, C.W., 2003. Extremes of some subsampled time series. J. Time Ser. Anal. 24, 579–590.
Turkman, K.F., 2011. Continuous and discrete time extremes of stationary processes. Research report, CEAUL 06/11, Center of Statistics and its Applications, University of Lisbon.
Turkman, K.F., Walker, A.M., 1984. On the asymptotic distributions of maxima of trigonometric polynomials with random coefficients. Adv. Appl. Probab. 16, 819–842.
Turkman, K.F., Walker, A.M., 1990. A stability result for the periodogram. Ann. Probab. 18, 1765–1783.
Walker, A.M., 1965. Some asymptotic results for the periodogram of a stationary time series. J. Aust. Math. Soc. 5, 107–128.
Zygmund, A., 1959. Trigonometric Series, vol. 2. Cambridge University Press, New York.

Part IX: Spectral and Wavelet Methods

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00021-1

The Estimation of Frequency

Barry G. Quinn
Department of Statistics, Macquarie University, NSW 2109, Australia

Abstract

Numerical methods have been used for fitting sinusoids to data since the middle of the 18th century. Since the discovery of the Fast Fourier Transform by Cooley and Tukey in 1965, the techniques for estimating frequency have become computationally feasible.

This review examines various techniques for estimating the frequency or frequencies of sinusoids in additive noise. The techniques fall into two categories – those based on Fourier, or frequency-domain methods, and those derived from a consideration of a small number of sample autocovariances. The Fourier techniques invariably have asymptotic variances of order T^{-3}, where T is the sample size, and are particularly useful when T is large and the signal is noisy, whereas the other techniques are usually statistically inefficient, with asymptotic variances of order T^{-1}, and are often biased, but because of their computational efficiency, can be useful when T is small and the signal is relatively noise free.

Keywords: frequency estimation, sinusoidal regression, periodogram, Fourier transform, resolution.

1. Introduction

An excellent historical account of the estimation of frequency, or the fitting of sinusoids, is contained in Bloomfield (1976), which also devotes a chapter to "Fitting Sinusoids." Further information is provided in Brillinger (1974, 1987), Heideman et al. (1984), and Priestley (1981).

Numerical methods have been used for fitting sinusoids since the middle of the 18th century. Of note is the method of Prony (1795) for fitting complex exponentials by solving systems of linear equations. The first technique that could be applied to other

than extremely small samples was the method of Buys-Ballot (1847) (Whittaker and Robinson, 1944). An implicit assumption in many of the techniques was that the sample size should be an integer multiple of the period – the data could then be lined up in such a way that addition would be coherent. Although there is some argument that the *Fast* Fourier transform algorithm predates the discovery by Cooley and Tukey (1965), it was not until then that computation of the discrete Fourier transform of large time series, and thus that frequency estimation based on Fourier methods, became feasible.

The history is slightly confused because of the use of Fourier methods in estimating both the frequency of a sinusoid or periodic function and the spectral density of a stationary stochastic process.

2. Basic model

Although we shall consider more general models, the basic noisy sinusoid satisfies an equation of the form

$$X_t = \mu + \rho\cos(\omega t + \phi) + \varepsilon_t \tag{1}$$

$$= \mu + \alpha\cos(\omega t) + \beta\sin(\omega t) + \varepsilon_t,\ t = 0, 1, \ldots, T-1 \tag{2}$$

where $\{\varepsilon_t\}$ is a zero-mean stochastic process with enough structure to ensure that parameter estimators have some decent asymptotic properties, and $\mu, \rho, \phi, \omega, \alpha = \rho\cos\phi, \beta = -\rho\sin\phi$ are unknown parameters. The form (2), is, for ω fixed, recognizable as a regression model with regressor variables $\cos(\omega t)$ and $\sin(\omega t)$. Although many papers have assumed, for example, that $\{\varepsilon_t\}$ is an independent and identically distributed sequence of random variables, or even normality, rarely are such assumptions needed. All we shall assume generally is that $\{\varepsilon_t\}$ is strictly stationary and ergodic, with continuous spectral density. We shall also assume the minimalist conditions of Section 2.5 in Quinn and Hannan (2001), referred to from now on as Q&H.

Figures 1 and 2 show a pure sinusoid, with $\mu = \beta = 0, \alpha = 1, \omega = 2\pi 35.5/128$, and $T = 128$, and a noisy version, which has the same parameters, but with pseudo-Gaussian white noise with variance 1. The noise standard deviation is thus the same as the amplitude of the sinusoid, making it virtually impossible to tell from Fig. 2 that there is a sinusoidal component present.

The most obvious method for estimating μ, α, β, and ω in (2) is regression. Since ω occurs nonlinearly, the least squares regression estimators may be defined by

$$\widehat{\omega}_T = \arg\min_{\omega} S_T(\omega)$$

$$S_T(\omega) = \min_{\mu,\alpha,\beta} \sum_{t=0}^{T-1} \{X_t - \mu - \alpha\cos(\omega t) - \beta\sin(\omega t)\}^2$$

$$\left(\widehat{\mu}_T, \widehat{\alpha}_T, \widehat{\beta}_T\right) = \arg\min_{\mu,\alpha,\beta} \sum_{t=0}^{T-1} \{X_t - \mu - \alpha\cos(\widehat{\omega}_T t) - \beta\sin(\widehat{\omega}_T t)\}^2 .$$

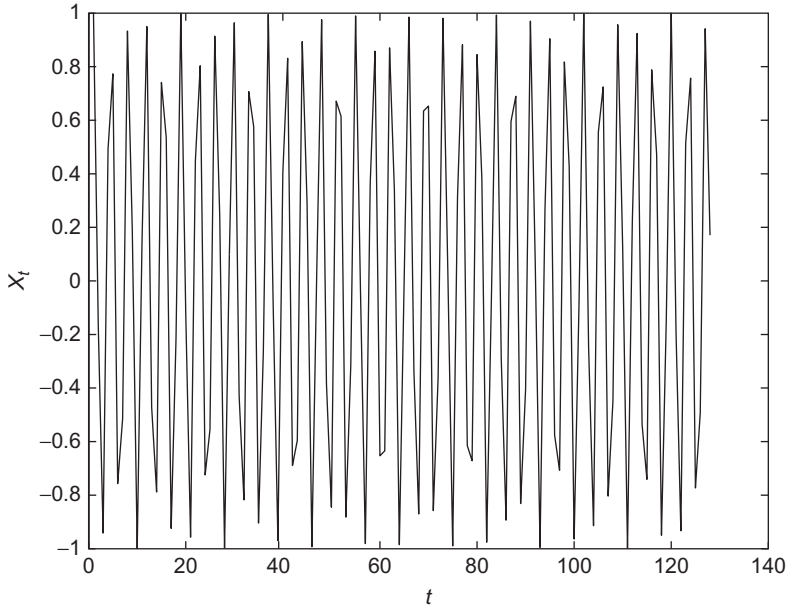

Fig. 1. Pure sinusoid.

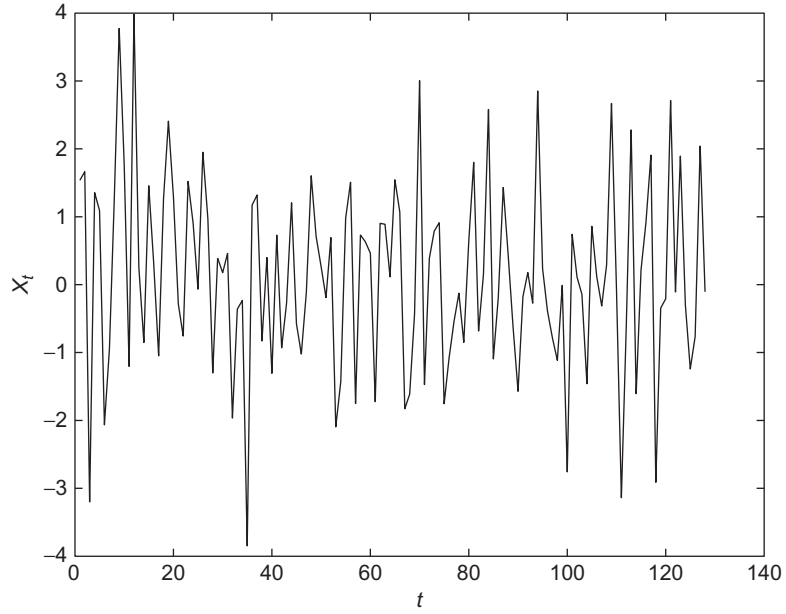

Fig. 2. Noisy sinusoid.

For fixed ω, let

$$M_T = \begin{bmatrix} 1 & 1 & 0 \\ 1 & \cos\omega & \sin\omega \\ \vdots & \vdots & \vdots \\ 1 & \cos\{\omega(T-1)\} & \sin\{\omega(T-1)\} \end{bmatrix}$$

be the design matrix for the regression equation given by (2). Then, using the identity

$$\sum_{t=0}^{T-1} e^{i\lambda t} = \begin{cases} T; & \lambda = 0, \bmod 2\pi \\ \frac{e^{i\lambda T}-1}{e^{i\lambda}-1}; & \lambda \neq 0, \bmod 2\pi, \end{cases}$$

we have, as $T \to \infty$, if $\omega \neq 0, \bmod \pi$,

$$M_T' M_T = \begin{bmatrix} T + O(1) & O(1) & O(1) \\ O(1) & \frac{T}{2} + O(1) & O(1) \\ O(1) & O(1) & \frac{T}{2} + O(1) \end{bmatrix}.$$

To see this, note for example that

$$\begin{aligned} \sum_{t=0}^{T-1} \cos^2(\omega t) &= \frac{1}{2}\sum_{t=0}^{T-1}\{1 + \cos(2\omega t)\} \\ &= \frac{T}{2} + \frac{1}{2}\,\mathrm{Re}\sum_{t=0}^{T-1} e^{i2\omega t} \\ &= \frac{T}{2} + \frac{1}{2}\,\mathrm{Re}\,\frac{e^{i2\omega T}-1}{e^{i2\omega}-1} \\ &= \frac{T}{2} + O(1). \end{aligned}$$

Thus, putting $\overline{X}_T = T^{-1}\sum_{t=0}^{T-1} X_t$, the regression sum of squares for fixed ω is

$$\begin{aligned}
&\sum_{t=0}^{T-1}\left(X_t - \overline{X}_T\right)^2 - S_T(\omega) \\
&\quad = \begin{bmatrix} \sum_{t=0}^{T-1} X_t \\ \sum_{t=0}^{T-1} X_t \cos(\omega t) \\ \sum_{t=0}^{T-1} X_t \sin(\omega t) \end{bmatrix}' \begin{bmatrix} T + O(1) & O(1) & O(1) \\ O(1) & \frac{T}{2} + O(1) & O(1) \\ O(1) & O(1) & \frac{T}{2} + O(1) \end{bmatrix}^{-1} \\
&\qquad \times \begin{bmatrix} \sum_{t=0}^{T-1} X_t \\ \sum_{t=0}^{T-1} X_t \cos(\omega t) \\ \sum_{t=0}^{T-1} X_t \sin(\omega t) \end{bmatrix} - T\overline{X}_T^2 \\
&\quad = I_X(\omega) + R_X(\omega),
\end{aligned}$$

where

$$I_X(\omega) = \frac{2}{T}\left|\sum_{t=0}^{T-1} X_t e^{-i\omega t}\right|^2, \tag{3}$$

and $R_X(\omega)$ may be shown to be smaller in order than $I_X(\omega)$, in the sense that, for any $\delta > 0$,

$$\sup_{\delta<\omega<\pi-\delta}\left|\frac{R_X(\omega)}{I_X(\omega)}\right| \to 0,$$

almost surely as $T \to \infty$.

The regression estimator of ω is thus asymptotically equivalent to the maximizer of the *periodogram* $I_X(\omega)$, which is a constant multiple of Schuster's (1898) periodogram and therefore has the same maximizer. Thus, the maximizer of the periodogram has the same asymptotic properties as the regression estimator of ω. In the same way, the estimator

$$\begin{bmatrix} \overline{X}_T \\ \frac{2}{T}\sum_{t=0}^{T-1} X_t \cos(\widehat{\omega}_T t) \\ \frac{2}{T}\sum_{t=0}^{T-1} X_t \sin(\widehat{\omega}_T t) \end{bmatrix}$$

of $[\mu\ \alpha\ \beta]'$ has the same asymptotic behavior as its regression estimator.

3. Properties of the periodogram maximizer

Because the periodogram maximizer is asymptotically equivalent to the least squares estimator, it follows that the asymptotic properties should mirror those of the maximum likelihood estimator constructed under Gaussian white noise assumptions, that is, under the assumption that the ε_t are normal, independent, and identically distributed. Since the information matrix, assuming that the ε_t have common variance σ^2, is for the parameter $\left[\mu\, \alpha\, \beta\, \omega\, \sigma^2\right]'$,

$$\frac{1}{\sigma^2}\begin{bmatrix} T & O(1) & O(1) & O(T) & 0 \\ O(1) & T/2+O(1) & O(1) & -\alpha T^2/4+O(T) & 0 \\ O(1) & O(1) & T/2+O(1) & \beta T^2/4+O(T) & 0 \\ O(T) & -\alpha T^2/4+O(T) & \beta T^2/4+O(T) & \left(\alpha^2+\beta^2\right)T^3/6+O(T^2) & 0 \\ 0 & 0 & 0 & 0 & \frac{T}{2\sigma^2} \end{bmatrix},$$

it follows that the Cramér-Rao lower bound for the variance of unbiased estimators of ω is

$$\frac{24\sigma^2}{\rho^2 T^3}\left\{1+O(1)\right\}.$$

The result is due to Whittle (1952), who also showed that under less restrictive conditions, $T^{3/2}\left(\widehat{\omega}_T-\omega\right)$ is asymptotically normally distributed with mean 0 and variance $24\sigma^2/\rho^2$. Walker (1971) proved rigorous results for the i.i.d. case. The definitive result is due to Hannan (1973), who showed that under very general colored noise assumptions, strong consistency and a central limit theorem hold, with $T^{3/2}\left(\widehat{\omega}_T-\omega\right)$ asymptotically normally distributed with mean 0 and variance $48\pi f(\omega)/\rho^2$, where

$$f(\lambda)=\frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma_j e^{-ij\lambda}$$

is the spectral density function of $\{\varepsilon_t\}$, and $\gamma_j=\operatorname{cov}\left(\varepsilon_t,\varepsilon_{t-j}\right)$. It is interesting to note that in the years between 1952 and 1974, the result was unknown in the engineering literature. In fact, in 1974, Rife and Boorstyn essentially derived the Cramér-Rao lower bound for the complex white Gaussian case, and it is their result that is still generally referred to in the engineering literature. The $T^{3/2}$ order in the asymptotic distribution may seem surprising at first, as the order is $T^{1/2}$, for example, for estimators of ARMA parameters. However, in time series regression, the order $T^{3/2}$ occurs, for example, in the central limit theorem for the slope parameter in a linear regression with time as the regressor.

Chen et al. (2000) have developed asymptotic theory for the estimators of frequency obtained by maximizing windowed periodograms.

4. Links with ARMA processes

The general solution of the differential equation

$$\frac{d^2x(t)}{dt^2} = -\omega^2 x(t)$$

is

$$x(t) = c_1 \cos(\omega t) + c_2 \sin(\omega t),$$

which is also the general solution of the autoregressive-like difference equation

$$x(t) - 2\cos\omega\, x(t-1) + x(t-2) = 0. \tag{4}$$

This was the starting point for the famous article of Yule (1927), where autoregressive processes were introduced to model periodicities evident in Wolfer's sunspot numbers. Yule imagined that although "errors of observation are practically eliminated" in the "departures of a simple harmonic pendulum from its position of rest," "unfortunately boys get into the room and start pelting the pendulum with peas, sometimes from one side and sometimes from the other." The displacement of the pendulum is then governed by the equation

$$x(t) - 2\cos\omega\, x(t-1) + x(t-2) = \varepsilon(t),$$

where the $\varepsilon(t)$ are perturbations or random errors. In Section 2 of his chapter, Yule proposed estimating $2\cos\omega$ by its least squares estimator

$$\frac{\sum_{t=1}^{T-1} \{x(t) + x(t-2)\}\, x(t-1)}{\sum_{s=1}^{T-1} x^2(s-1)}. \tag{5}$$

Another obvious approach is to estimate $2\cos\omega$ by the Yule–Walker estimator (see, e.g., Priestley (1981))

$$2\frac{\sum_{t=1}^{T-1} x(t)x(t-1)}{\sum_{s=0}^{T-1} x^2(s)}.$$

In Section 3 of his paper, Yule proposed fitting the AR(2) model

$$x(t) - b_1 x(t-1) + b_2 x(t-2) = \varepsilon(t)$$

to the sunspot numbers. Since the zeros of $z^2 - 2r\cos\omega z + r^2$ are $re^{\pm i\omega}$ and the general solution of

$$x(t) - 2r\cos\omega\, x(t-1) + r^2 x(t-2) = 0$$

is

$$x(t) = c_1 r^t \cos(\omega t) + c_2 r^t \sin(\omega t),$$

Yule proposed estimating the frequency as the solution $\widehat{\omega}$ in $(0, \pi)$ of

$$z^2 - \widehat{b}_1 z + \widehat{b}_2 = 0$$
$$z = \widehat{r}e^{\pm i\widehat{\omega}},$$

where $\widehat{b}_1$ and $\widehat{b}_2$ are the estimators of b_1 and b_2, respectively, constructed from sample autocorrelations (the *Yule–Walker* relations) and assuming that $\widehat{b}_1^2 < 4\widehat{b}_2$.

Using data from the same years available to Yule (1749–1924), the estimates of the period $P = 2\pi/\omega$ using the raw data and the Yule–Walker method are 9.97 years for the Section 2 and 10.55 years for the Section 3 techniques, respectively. As many other authors have transformed the sunspot numbers by taking square roots, we applied the techniques to the square roots of the series. The period estimates were then 10.26 and 10.85, respectively.

5. Autoregressive approximation

In light of Yule's work, many authors have explored the use of autoregressive approximation to estimate frequency. Suppose $\{X(t)\}$ is a second-order stationary autoregressive process that satisfies the equation

$$X(t) + \sum_{j=1}^{p} \beta_j X(t-j) = \varepsilon(t), \tag{6}$$

where $\{\varepsilon(t)\}$ is uncorrelated and second-order stationary. Its spectral density function is then

$$f(\lambda) = \frac{\sigma^2}{2\pi}\left|1 + \sum_{j=1}^{p} \beta_j e^{ij\lambda}\right|^{-2}. \tag{7}$$

If the polynomial $\beta(z) = 1 + \sum_{j=1}^{p} \beta_j z^j$ has a complex pair of zeros, say $r^{-1}e^{\pm i\omega}$, where $0 < r < 1$, then the homogeneous difference equation

$$X(t) + \sum_{j=1}^{p} \beta_j X(t-j) = 0,\ t = 0, 1, \ldots$$

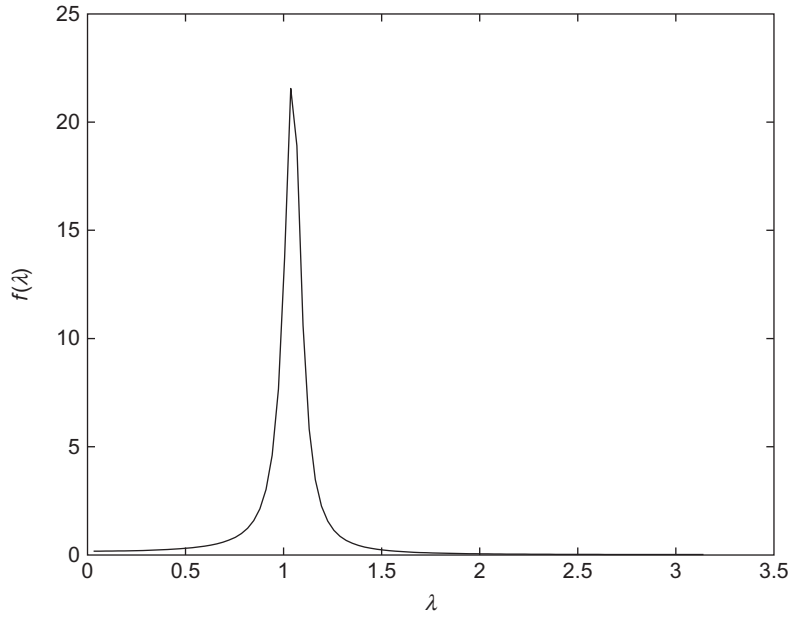

Fig. 3. Spectral density function of an AR(2) process.

will have a solution of the form

$$cr^t \cos(\omega t + \phi).$$

If r is close to 1, and the other zeros of $\beta(z)$ have moduli much greater than r^{-1}, then $f(\lambda)$ will have a peak close to ω. For example, suppose $p = 2, \beta_1 = -1.9\cos(\pi/3)$, $\beta_2 = 0.95^2$, and $\sigma^2 = 1$. Then

$$\beta(z) = \left(1 - 0.95e^{-i\pi/3}z\right)\left(1 - 0.95e^{i\pi/3}z\right)$$

has zeros $0.95^{-1}e^{\pm i\pi/3}$. The spectral density is depicted in Fig. 3.

The peak, however, is *not* at $\pi/3 \sim 1.0472$, but at

$$\arccos\left(\frac{\cos(\pi/3)\left(1 + 0.95^2\right)}{2 \times 0.95}\right) \sim 1.0464.$$

If $\beta_1 = -1.98\cos(\pi/3)$ and $\beta_2 = 0.99^2$, the peak is now at

$$\arccos\left(\frac{\cos(\pi/3)\left(1 + 0.99^2\right)}{2 \times 0.99}\right) \sim 1.0472,$$

illustrating what happens as the zeros of $\beta(z)$ approach the unit circle. It follows that the estimated spectral density, formed from (7), but using estimated parameters, has maximizer that is *not* a consistent estimator of ω.

Suppose that $\{X(t)\}$ satisfies (2), and that an autoregression of some order is fitted using the Yule–Walker relations. For $j = 0, 1, \ldots$, let C_j denote the jth sample autocovariance given by

$$C_j = T^{-1} \sum_{t=j}^{T-1} \left(X(t) - \overline{X}_T\right)\left(X(t-j) - \overline{X}_T\right),$$

where

$$\overline{X}_T = T^{-1} \sum_{t=0}^{T-1} X_t.$$

Then

$$\begin{aligned} \overline{X}_T &\to \mu, \\ C_j &\to \gamma_j + \rho^2 \cos(\omega j)/2, \end{aligned}$$

almost surely as $T \to \infty$, where $\gamma_j = \operatorname{cov}\left(\varepsilon_t, \varepsilon_{t-j}\right)$. It is therefore impossible to estimate ω using the C_j, without assuming something about the γ_j. If, for example, we assume that $\{\varepsilon(t)\}$ is white, then we may estimate ω consistently from the two equations

$$\begin{aligned} C_1 &= \frac{1}{2}\rho^2 \cos\widehat{\omega} \\ C_2 &= \frac{1}{2}\rho^2 \cos(2\widehat{\omega}), \end{aligned} \tag{8}$$

or the equation

$$\frac{C_2}{C_1} = \frac{\cos(2\widehat{\omega})}{\cos\widehat{\omega}} = \frac{2\cos^2\widehat{\omega} - 1}{\cos\widehat{\omega}}. \tag{9}$$

The only (strongly) consistent estimator of ω constructed from C_1 and C_2 is thus

$$\widehat{\omega} = \arccos\left(\frac{C_2 + \sqrt{C_2^2 + 8C_1^2}}{4C_1}\right). \tag{10}$$

This estimator should be contrasted with Yule's two estimators

$$\arccos\left(\frac{C_1}{C_0}\right) \tag{11}$$

and

$$\arccos\left(-\frac{b_1}{2\sqrt{b_2}}\right), \tag{12}$$

where

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} C_0 & C_1 \\ C_1 & C_0 \end{bmatrix}^{-1} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix},$$

both of which are consistent only if $\sigma^2 = 0$, that is, if there is no noise, in which case the estimators are exact, and thus trivially consistent. For the raw sunspot numbers, the estimates of period using (10), (11), and (12) are 11.37, 9.94, and 10.55 years, respectively. The transformed series yields estimates 11.62, 10.21, and 10.85 years. The Quinn and Fernandes (1991) estimates and the periodogram maximizer are 11.38 and 11.36 years, respectively, for the raw data and 11.34 and 11.33 for the transformed series.

The differences between the above estimates are noteworthy and reflect the different assumptions made. The periodogram maximizer, Quinn–Fernandes, and autoregressive estimator have been derived for the case of a sinusoid in additive noise, whereas Yule's two estimators assume the underlying process to be the solution of a difference equation with stochastic forcing term.

If the process is indeed a noisy sinusoid, then the autoregression-based techniques are clearly inappropriate. For even if (10) is used to estimate ω, and $\{\varepsilon(t)\}$ is i.i.d. with common variance σ^2, it is easily shown that

$$T^{\frac{1}{2}} \begin{bmatrix} C_1 - \frac{\rho^2}{2}\cos\omega \\ C_2 - \frac{\rho^2}{2}\cos(2\omega) \end{bmatrix}$$

is asymptotically normally distributed with mean 0 and covariance matrix

$$\left(\sigma^2\right)^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + 2\rho^2\sigma^2 \begin{bmatrix} \cos\omega \\ \cos(2\omega) \end{bmatrix} \begin{bmatrix} \cos\omega & \cos(2\omega) \end{bmatrix},$$

and consequently, the estimator $\widehat{\omega}$ of ω is such that $T^{\frac{1}{2}}(\widehat{\omega} - \omega)$ is asymptotically normal with mean 0 and variance

$$2\left(\frac{\sigma^2}{\rho^2}\right)^2 \frac{\cos^2\omega + \cos^2(2\omega)}{\left(2\cos^2\omega + 1\right)^2}.$$

If a higher-order autoregression is fitted, the maximizer of the estimated spectral density function will still not be a consistent estimator of ω. It has been conjectured that if an autoregression is fitted, for example, using AIC, then the estimated order will be an increasing function of T and the asymptotic "bias" converges to 0. However, this has not been proven.

It is easy to see why the fitting of a fixed-order autoregression, with order, say, p, can never produce an estimator that has as good asymptotic performance as the periodogram maximizer. Let $\widehat{\beta}$ be the vector of autoregressive estimators. Again assume that $\{\varepsilon(t)\}$ is i.i.d. Then

$$\widehat{\omega} = h(\widehat{\beta})$$
$$\widehat{\beta} = g(C),$$

where $C = \begin{bmatrix} C_0 & \cdots & C_p \end{bmatrix}$ and g and h are differentiable. Since

$$\begin{bmatrix} C_0 - \sigma^2 - \frac{\rho^2}{2} \\ C_1 - \frac{\rho^2}{2}\cos\omega \\ \vdots \\ C_p - \frac{\rho^2}{2}\cos(p\omega) \end{bmatrix}$$

converges almost surely to 0, and

$$T^{\frac{1}{2}} \begin{bmatrix} C_0 - \sigma^2 - \frac{\rho^2}{2} \\ C_1 - \frac{\rho^2}{2}\cos\omega \\ \vdots \\ C_p - \frac{\rho^2}{2}\cos(p\omega) \end{bmatrix}$$

is asymptotically normally distributed with some covariance matrix Σ, it follows assuming that

$$\omega = h(g(\gamma)),$$

where

$$\gamma = \begin{bmatrix} \sigma^2 + \frac{\rho^2}{2} \\ \frac{\rho^2}{2}\cos\omega \\ \vdots \\ \frac{\rho^2}{2}\cos(p\omega) \end{bmatrix},$$

that $\widehat{\omega}$ is consistent and $T^{1/2}(\widehat{\omega} - \omega)$ is asymptotically normal. It is therefore the case that the rate of convergence is much less than that of the periodogram maximizer. Moreover, the estimator will be consistent only if the noise is white, whereas the periodogram maximizer has excellent properties even when the noise is colored.

6. Pisarenko's technique

Because of the use of the sample autocovariance matrix in the estimation of autoregressive parameters, there has been much interest in the use of sample autocovariance matrices in the signal processing literature in the estimation of frequency. Pisarenko's (1973) procedure is popular: let

$$\mathbb{C}_p = \begin{bmatrix} C_0 & C_1 & \cdots & C_p \\ C_1 & C_0 & \cdots & C_{p-1} \\ \vdots & \ddots & \ddots & \vdots \\ C_p & \cdots & C_1 & C_0 \end{bmatrix},$$

and let $d = \begin{bmatrix} d_0 & d_1 & d_2 \end{bmatrix}'$ denote an eigenvector corresponding to the smallest eigenvector of $\mathbb{C}_2$. ω is estimated by that argument of the zeros of $d_0 + d_1 z + d_2 z^2$, which is in $(0, \pi)$. The motivation for the technique is that under white noise conditions,

$$\mathbb{C}_2 \underset{a.s.}{\to} \sigma^2 I_3 + \frac{\rho^2}{2} \begin{bmatrix} 1 & \cos\omega & \cos(2\omega) \\ \cos\omega & 1 & \cos\omega \\ \cos(2\omega) & \cos\omega & 1 \end{bmatrix}$$
$$= \sigma^2 I_3 + \frac{\rho^2}{2} \left\{ cc' + ss' \right\},$$

where

$$c' = \begin{bmatrix} 1 & \cos\omega & \cos(2\omega) \end{bmatrix}$$
$$s' = \begin{bmatrix} 0 & \sin\omega & \sin(2\omega) \end{bmatrix}.$$

Now

$$\sigma^2 I_3 + \frac{\rho^2}{2} \left\{ cc' + ss' \right\}$$

has smallest eigenvalue σ^2, and, since

$$\begin{bmatrix} 1 & -2\cos\omega & 1 \end{bmatrix}'$$

is orthogonal to both c and s, it is a left eigenvector corresponding to the smallest eigenvalue of the almost sure limit of $\mathbb{C}_2$. As the zeros of $1 - 2z\cos\omega + z^2$ are $e^{\pm i\omega}$, it follows that the Pisarenko estimator is strongly consistent.

Because of the Toeplitz structure of $\mathbb{C}_2$, its left eigenvectors are multiples of the forms $\begin{bmatrix} 1 & a & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & 0 & -1 \end{bmatrix}$. The equation

$$\begin{bmatrix} C_0 & C_1 & C_2 \\ C_1 & C_0 & C_1 \\ C_2 & C_1 & C_0 \end{bmatrix} \begin{bmatrix} 1 \\ a \\ 1 \end{bmatrix} = \lambda \begin{bmatrix} 1 \\ a \\ 1 \end{bmatrix}$$

has solutions

$$a = \frac{-C_2 + \sqrt{C_2^2 + 8C_1^2}}{2C_1}, \quad \lambda = C_0 + \frac{C_2 + \sqrt{C_2^2 + 8C_1^2}}{2},$$
$$a = \frac{-C_2 - \sqrt{C_2^2 + 8C_1^2}}{2C_1}, \quad \lambda = C_0 + \frac{C_2 - \sqrt{C_2^2 + 8C_1^2}}{2},$$

whereas

$$\begin{bmatrix} C_0 & C_1 & C_2 \\ C_1 & C_0 & C_1 \\ C_2 & C_1 & C_0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \lambda \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

has solution

$$\lambda = C_0 - C_2.$$

The minimum eigenvalue is therefore the minimum of

$$C_0 + \frac{C_2 - \sqrt{C_2^2 + 8C_1^2}}{2}$$

and

$$C_0 - C_2.$$

Now

$$C_0 - C_2 - C_0 - \frac{C_2 - \sqrt{C_2^2 + 8C_1^2}}{2} = \frac{-3C_2 + \sqrt{C_2^2 + 8C_1^2}}{2},$$

which converges almost surely to

$$\frac{\sqrt{\cos^2(2\omega) + 8\cos^2\omega} - 3\cos(2\omega)}{2} = 1 - \cos(2\omega) > 0.$$

Thus, almost surely as $T \to \infty$, the smallest eigenvalue is

$$C_0 + \frac{C_2 - \sqrt{C_2^2 + 8C_1^2}}{2}.$$

Pisarenko's estimator is therefore the positive arg of the zeros of

$$1 + \frac{-C_2 - \sqrt{C_2^2 + 8C_1^2}}{2C_1} z + z^2,$$

which occur as a complex conjugate pair with modulus 1. Denoting Pisarenko's estimator by $\widehat{\omega}_P$, it follows that

$$-2\cos\widehat{\omega}_P = \frac{-C_2 - \sqrt{C_2^2 + 8C_1^2}}{2C_1},$$

or

$$\widehat{\omega}_P = \arccos \frac{C_2 + \sqrt{C_2^2 + 8C_1^2}}{4C_1},$$

which is the same as (10), not surprisingly, as we have already shown that this is the *only* strongly consistent estimator constructed from C_0, C_1, and C_2 and that Pisarenko's estimator is strongly consistent.

Pisarenko's technique generalizes to estimating the frequencies of a noisy sum of p sinusoids. In this case, the sample autocovariance matrix constructed from $C_0, \ldots, C_{2p}$ is constructed, the eigenvector corresponding to the smallest eigenvector again computed, and the positive arguments of the zeros of the polynomial constructed from the elements of the eigenvector are used as estimators of the p frequencies. For details, see Pisarenko (1973), and for an asymptotic analysis, see Sakai (1984).

7. MUSIC

The MUltiple SIgnal Characterization technique of Schmidt (1981, 1986) comes from the array processing literature, providing estimators of direction of arrival rather than frequency. The problems are very similar, however, and MUSIC has become a popular technique for estimating frequency. As with many of the previous techniques, the noise must be white for the technique to work.

Let $\left\{\widehat{P}_j; j = 1, \ldots, K\right\}$ denote normalized eigenvectors (i.e., $\widehat{P}_j' \widehat{P}_j = 1$ for each j) corresponding to the decreasing eigenvalues of the sample autocovariance matrix $\mathbb{C}_K$, where $K \geq 3$. The MUSIC estimator of ω is defined to be the minimizer of

$$\sum_{k=3}^{K} \left|\widehat{P}_k' e_K(\omega)\right|^2,$$

where, letting z^* be the complex conjugate transpose of z,

$$e_K^*(\omega) = \begin{bmatrix} 1 & e^{-i\omega} & \cdots & e^{-iK\omega} \end{bmatrix}.$$

Since

$$\begin{aligned} \sum_{k=1}^{K} \left|\widehat{P}_k' e_K(\omega)\right|^2 &= e_K^*(\omega) \sum_{k=1}^{K} \widehat{P}_k \widehat{P}_k' e_K(\omega) \\ &= K + 1, \end{aligned}$$

it follows that the MUSIC estimator of ω is the maximizer of what is termed the MUSIC spectrum

$$\sum_{k=1}^{2} \left|\widehat{P}_k' e_K(\omega)\right|^2.$$

In the special case where $K = 3$, the MUSIC estimator is (asymptotically) the same as Pisarenko's, since the minimizer of

$$\left|1 + \frac{-C_2 - \sqrt{C_2^2 + 8C_1^2}}{2C_1} e^{i\omega} + e^{i2\omega}\right|^2$$

satisfies the equation

$$2\cos\omega = -\frac{-C_2 - \sqrt{C_2^2 + 8C_1^2}}{2C_1}.$$

For general K and the more general case of more than one frequency, the analysis is contained in Q&H, where the results for complex sinusoids are also presented.

A related estimator is the minimizer of

$$\sum_{k=3}^{K} \widehat{\lambda}_k^{\alpha} \left|\widehat{P}_k' e_K(\omega)\right|^2,$$

where the $\widehat{\lambda}_k$ are the eigenvalues of $\mathbb{C}_K$, in decreasing order, and $\alpha \in \mathbb{R}$. When $\alpha = -1$, the technique is known as the EV method (Johnson, 1982). It is shown in Q&H that the asymptotic properties are the same as those of MUSIC, for any α. In particular, the asymptotic variance of the estimator is $O\left(T^{-1}\right)$.

8. An efficient technique based on ARMA filtering

In light of the fact that

$$X_t - 2\cos\omega X_{t-1} + X_{t-2} = \varepsilon_t - 2\cos\omega\, \varepsilon_{t-1} + \varepsilon_{t-2},$$

it makes sense to consider ARMA(2, 2) estimation techniques that constrain the zeros of the autoregressive polynomial to be near, or even on, the unit circle. Nehorai and Porat (1986) and Fernandes et al. (1987) suggested iterative estimation procedures with zeros approaching the unit circle with each iteration. Li and Kedem (1993) have developed a similar technique, but bounded the zeros away from the unit circle. Techniques that bound the zeros away from the unit circle inevitably have asymptotic variances of order $O\left(T^{-1}\right)$. However, although the statistical properties of estimation techniques with zeros approaching the unit circle are unknown, simulations suggested that they were better than Pisarenko's estimator and MUSIC. The fact that constraining the autoregressive polynomial's zeros to be *on* the unit circle is anathema to Engineers is probably the reason why the following idea had not been tried.

Suppose we fit the ARMA(2,2) model

$$X_t + \beta X_{t-1} + X_{t-2} = \varepsilon_t + \alpha\varepsilon_{t-1} + \varepsilon_{t-2} \tag{13}$$

in the following way:

1. Let $\widehat{\alpha}_0$ be an initial value of α. (This should therefore be an estimator of $-2\cos\omega$.)
2. Put

$$\xi_t = X_t - \widehat{\alpha}_0 \xi_{t-1} - \xi_{t-2}$$
$$\xi_{-1} = \xi_{-2} = 0.$$

   Then

$$\varepsilon_t = \xi_t + \beta \xi_{t-1} + \xi_{t-2},$$

   and so, the minimizer with respect to β of

$$\sum_{t=0}^{T-1} \varepsilon_t^2 = \sum_{t=0}^{T-1} (\xi_t + \beta \xi_{t-1} + \xi_{t-2})^2$$

   is found by regressing $\xi_t + \xi_{t-2}$ on $-\xi_{t-1}$. We would thus estimate β by

$$-\frac{\sum_{t=0}^{T-1} (\xi_t + \xi_{t-2})\, \xi_{t-1}}{\sum_{s=0}^{T-1} \xi_{s-1}^2}.$$

3. Since α and β should be equal, we could then replace $\widehat{\alpha}_0$ by

$$\begin{aligned}\widehat{\alpha}_1 &= -\frac{\sum_{t=0}^{T-1} (\xi_t + \xi_{t-2})\, \xi_{t-1}}{\sum_{s=0}^{T-1} \xi_{s-1}^2} \\ &= -\frac{\sum_{t=0}^{T-1} (X_t - \widehat{\alpha}_0 \xi_{t-1})\, \xi_{t-1}}{\sum_{s=0}^{T-1} \xi_{s-1}^2} \\ &= \widehat{\alpha}_0 - \frac{\sum_{t=0}^{T-1} X_t \xi_{t-1}}{\sum_{s=0}^{T-1} \xi_{s-1}^2}\end{aligned}$$

   and carry out step 2 again, repeating until the process "converges."

There is no guarantee *a priori* that this might produce any sensible estimation procedure. However, simple simulations suggested that the increment

$$-\frac{\sum_{t=0}^{T-1} X_t \xi_{t-1}}{\sum_{s=0}^{T-1} \xi_{s-1}^2}$$

should be doubled to improve convergence. The procedure introduced in Quinn and Fernandes is

1. Let $\widehat{\alpha}_0$ be an initial value of $\alpha = -2\cos\omega$.

2. For $j = 0, 1, \ldots,$
   (a) Put, for $t = 0, \ldots, T$,

$$\xi_t = X_t - \widehat{\alpha}_j \xi_{t-1} - \xi_{t-2}$$
$$\xi_{-1} = \xi_{-2} = 0.$$

   (b) Let

$$\widehat{\alpha}_{j+1} = \widehat{\alpha}_j - 2\frac{\sum_{t=0}^{T-1} X_t \xi_{t-1}}{\sum_{s=0}^{T-1} \xi_{s-1}^2}$$

   (c) Repeat step 2 unless $\left|\sum_{t=0}^{T-1} X_t \xi_{t-1} / \sum_{s=0}^{T-1} \xi_{s-1}^2\right|$ is acceptably small.
3. With $\widehat{\alpha}$ the current value of $\widehat{\alpha}_{j+1}$, put

$$\widehat{\omega}_{QF} = \arccos\left(-\frac{\widehat{\alpha}}{2}\right).$$

Note that the procedure involves only a number of filters and simple arithmetic operations, which can be carried out frugally using mathematical and statistical packages. The question of starting values and "convergence" criterion has a simple answer. Under the same conditions as assumed for the periodogram maximizer $\widehat{\omega}_P$, if the initial estimator of ω is accurate to order $o_P\left(T^{-1/2}\right)$, the estimator has the same asymptotics as the periodogram maximizer. Moreover, if

$$\widehat{\alpha}_0 = -2\cos\left(2\pi k_T / T\right),$$

where

$$k_T = \arg \max_{1 \le j \le T/2} I_X\left(\frac{2\pi j}{T}\right), \tag{14}$$

it may be shown that

$$T^{3/2}\left(\widehat{\omega}_{QF} - \widehat{\omega}_P\right) \to 0 \tag{15}$$

in probability as $T \to \infty$, where

$$\widehat{\omega}_{QF} = \arccos\left(-\frac{\widehat{\alpha}_2}{2}\right).$$

In other words, to estimate ω with the same efficiency as the periodogram maximizer, the iterations above need to be carried out only twice. In particular, although the procedure was motivated by least squares, $\{\varepsilon_t\}$ need not be white. In light of (15), the two estimators have the same central limit theorem.

Note: Truong-Van (1990) proposed an estimator based on finding zeros of $\sum_{t=0}^{T-1} X_t \xi_{t-1}$. The procedure was motivated by the fact that the solution to the difference equation

$$\xi_t - 2\cos\lambda \xi_{t-1} + \xi_{t-2} = \cos\left(\omega t + \phi\right)$$

"rings" when $\lambda = \omega$, that is of the form

$$\xi_t = ct \cos(\omega t + \nu).$$

Truong-Van termed the approach "Amplified Harmonics." Since the Quinn–Fernandes technique also produces zeros of $\sum_{t=0}^{T-1} X_t \xi_{t-1}$, the estimators theoretically share the same asymptotic behavior. Although Truong-Van assumed the noise process $\{\varepsilon_t\}$ to be ARMA, this is not necessary.

Note: The zeros of $\sum_{t=0}^{T-1} X_t \xi_{t-1}$ are also the zeros of

$$\sum_{j=1}^{T-1} \sin(j\lambda)\, C_j,$$

which, because of the relation

$$C_j = \frac{1}{4\pi} \int_{-\pi}^{\pi} e^{ij\lambda} I_X(\lambda) d\lambda,$$

are local maximizers of

$$\begin{aligned} \kappa_X(\lambda) &= \int_{-\pi}^{\pi} I_X(\gamma) \mu_T(\lambda - \gamma)\, d\gamma \\ &= \int_{-\pi}^{\pi} I_X(\gamma) \mu(\lambda - \gamma)\, d\gamma, \end{aligned}$$

where

$$\begin{aligned} \mu_T(\lambda) &= \sum_{k=1}^{T-1} \frac{\cos(k\lambda)}{k}, \\ \mu(\lambda) &= \sum_{k=1}^{\infty} \frac{\cos(k\lambda)}{k} \\ &= -\frac{1}{2} \log\left\{4 \sin^2\left(\frac{\lambda}{2}\right)\right\}. \end{aligned}$$

The function $\kappa_X(\lambda)$ is therefore a smoothed version of the periodogram – the convolution of the periodogram with $\mu_T(\lambda)$, which is $\sim \log T$ at $\lambda = 0$, but which converges for any fixed λ. Further details are given in Q&H. Figure 4 depicts μ_T, for $T = 128, 256, 384, \ldots, 1024$ and μ.

Finally, we note that Song and Li (2000) contains an asymptotic analysis of the technique of Li and Kedem (1993) , modified to allow the zeros of the autoregressive polynomial to converge to the unit circle at a rate depending on T. They claim

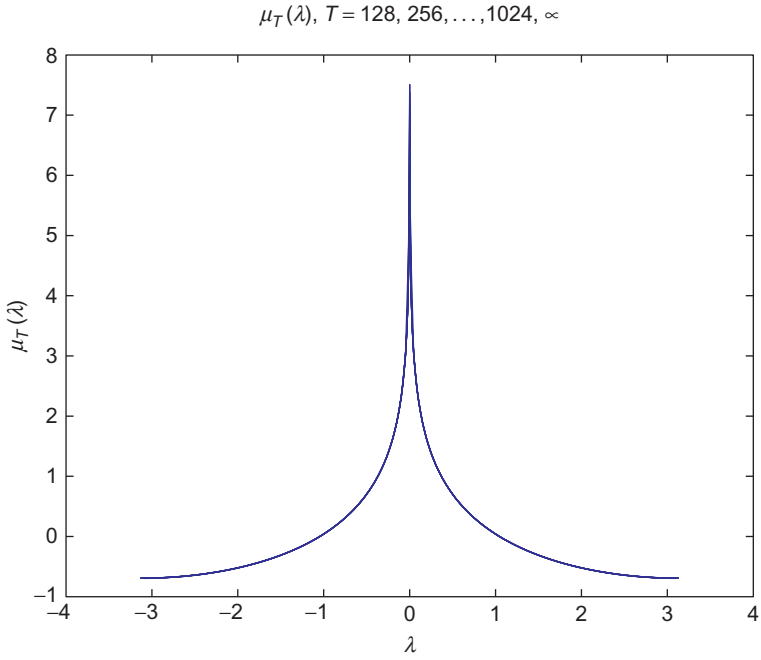

Fig. 4. The kernel μ_T.

that their technique allows any initial value of ω and still produces an estimator with order of asymptotic variance arbitrarily close to T^{-3}. Their claim that the Quinn and Fernandes technique requires an initial estimator accurate to order $o(T^{-1})$ is incorrect, as indicated above.

9. Maximizing the periodogram: practicalities

Figures 5 and 6 show the periodogram and $\kappa_X(\lambda)$ for the same time series, which has been simulated from (1) with $\rho = 1, \omega = 2\pi 35.3/1024, \phi = 0, T = 1024$, and $\{\varepsilon_t\}$ Gaussian and white with variance 1.

Although it may appear that the periodogram is better at "detecting" the sinusoid, it should be obvious from Figs 5 and 6 that it will be easier numerically to find the maximizer of κ_T than the periodogram maximizer, since the derivative of the periodogram near the main "spike" is changing very quickly locally. In fact, Newton's method, applied to find a zero of $I'_X(\omega)$, with initial estimator $2\pi k_T/T$, where k_T is given by (14), is not guaranteed to produce the periodogram maximizer. This is because of the fact that "sidelobes" of the periodogram occur within $O(T^{-1})$ of the true frequency, while κ_T does not have any sidelobes nearby. For details, see Rice and Rosenblatt (1988) and Quinn et al. (2008). The latter show that if

$$n_T = \arg \max_{1 \le j \le 2T} I_X\left(\frac{2\pi j}{4T}\right),$$

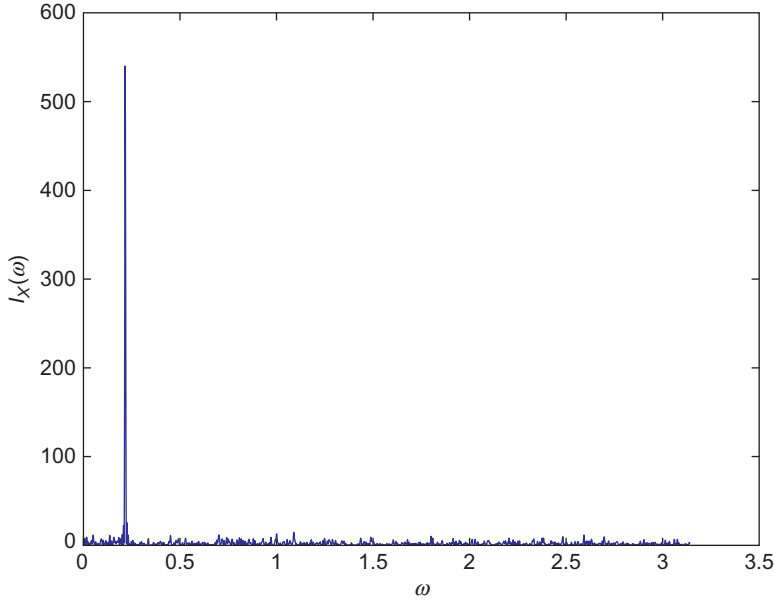

Fig. 5. Periodogram of noisy sinusoid.

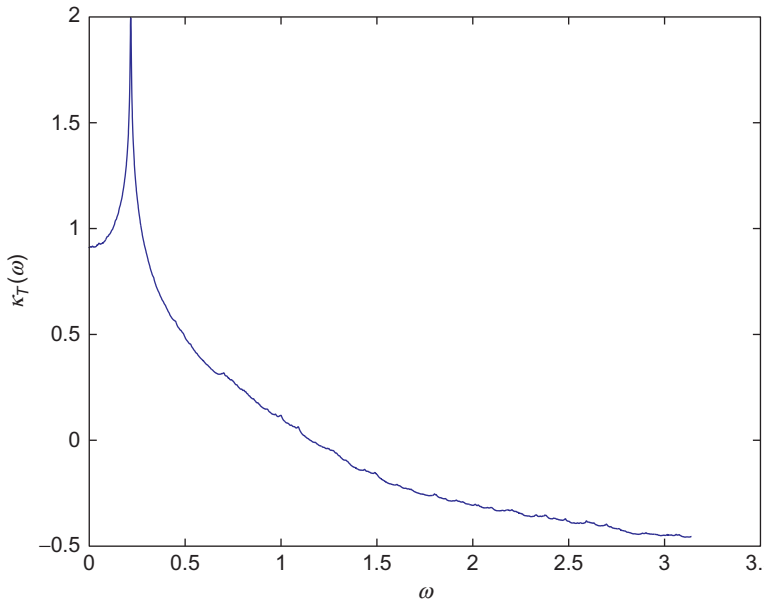

Fig. 6. κ_T for same time series.

obtained, for example, by zero-padding, that is by Fourier-transforming

$$\{X_0, \ldots, X_{T-1}\}$$

with $3T$ zeros appended, then Newton's method can be shown to work if started with the estimator $2\pi n_T/(4T)$. It is also shown that Newton's method works with initial estimator $2\pi k_T/T$ if applied to find the maximizer of $\{I_X(\omega)\}^\alpha$, with $\alpha \leq 0.373$.

10. Discrete Fourier transform-based methods

Given the problems associated with maximizing $I_X(\omega)$ as a continuous function of frequency, and the ease of computing $I_X(\omega)$ at the Fourier frequencies $\{\omega_j = 2\pi j/T; 0 \leq j < T\}$, it makes sense to consider estimation using the Fourier coefficients

$$Y_j = \sum_{t=0}^{T-1} X_t e^{-i\omega_j t}. \tag{16}$$

If

$$\omega = \frac{2\pi\,(k + \delta_T)}{T}, \tag{17}$$

where $\delta_T = O(1)$, then

$$Y_{k+j} = \frac{T\rho e^{i\phi}}{2} \frac{e^{i2\pi\delta_T} - 1}{2\pi i\,(\delta_T - j)} + \sum_{t=0}^{T-1} \varepsilon_t e^{-i\omega_j t} + O(1). \tag{18}$$

Bartlett (1967) appears to has been the first to use this fact to estimate δ_T and thus ω by minimizing with respect to δ_T and the complex constant D

$$\sum_j \left| Y_{k+j} - \frac{D}{\delta_T - j} \right|^2, \tag{19}$$

the sum being over a small number of j close to 0. Bartlett did this by minimizing (19) with respect to D for δ_T on some grid and then by interpolation. The asymptotic behavior is described in Q&H. In modern times, minimization of (19) is not as problematic. A closed-form expression is preferable, however, in real-time systems, where data are processed online, rather than offline.

From (18), we have

$$\frac{Y_{k+j}}{Y_j} = \frac{\delta_T}{\delta_T - j} + o(1),$$

almost surely as $T \to \infty$, since each of $T^{-1}\sum_{t=0}^{T-1} \varepsilon_t e^{-i\omega_j t}$ converges almost surely to 0. Each of the equations

$$\operatorname{Re}\frac{Y_{k+j}}{Y_k} = \frac{\delta_T}{\delta_T - j}, \quad j = \pm 1, \pm 2, \ldots,$$

thus provides a "strongly consistent" estimator of ω via (17). For example, with $j = 1$,

$$\widehat{\delta}_T = \frac{\operatorname{Re}\left(Y_{k+j}/Y_j\right)}{\operatorname{Re}\left(Y_{k+j}/Y_j\right) - 1}.$$

What remains is to define j and k in a sensible way, so that asymptotics also make sense. One such algorithm, the Fourier Transform Interpolator (FTI), is given in Quinn (1992, 1994):

Algorithm 1. (FTI)

1. Let k_T be given by (14).
2. Let $\widehat{\delta}_{jT} = \frac{jR_j}{R_j - 1}$, where $R_j = \frac{\operatorname{Re}\left(Y_{k+j}/Y_j\right)}{\operatorname{Re}\left(Y_{k+j}/Y_j\right) - 1}$.
3. Let $\widehat{\delta}_T = \widehat{\delta}_{1T}$. If $\widehat{\delta}_{jT} > 0$, $j = \pm 1$, let $\widehat{\delta}_T = \widehat{\delta}_{-1T}$.
4. Put $\widehat{\omega}_T = 2\pi\left(k_T + \widehat{\delta}_T\right)/T$.

It is shown in Quinn (1994) and Quinn and Hannan (2001) that $T^{3/2}(\log T)^{-1/2-\nu}(\widehat{\omega}_T - \omega)$ converges almost surely to 0, for all $\nu > 0$, and that the distribution function of

$$T^{3/2} v_T^{-1}(\widehat{\omega}_T - \omega)$$

converges to that of the standard normal, where

$$v_T^2 = \frac{16\pi^3 f(\omega)}{\rho^2} \frac{\pi^2 \delta_T^2}{\sin^2(\pi\delta_T)} (1 - |\delta_T|)^2 \left\{(1 - |\delta_T|)^2 + \delta_T^2\right\}$$

and $\delta_T = T\omega/(2\pi)$ minus its nearest integer, and is therefore in $\left[-\frac{1}{2}, \frac{1}{2}\right]$. That the "asymptotic variance" should depend on T is not surprising, as the Fourier frequencies change with T. The choice between two possible estimators in step 3 of FTI is motivated in Quinn (1992, 1994) and Quinn and Hannan (2001). MacLeod (1998) has pointed out that putting $\widehat{\delta}_T = \widehat{\delta}_{-1T}$ if $R_{-1} > R_1$ is better, especially when the closest integer to $T\omega/(2\pi)$ is known a priori. Although the procedure uses only three Fourier coefficients, the ratio

$$\frac{v_T^2}{48\pi f(\omega)/\rho^2},$$

the asymptotic efficiency relative to the periodogram maximizer, is largest $\left(\pi^2/3 \sim 3.2899\right)$ when $\delta_T = 0$, and smallest $\left(\pi^4/96 \sim 1.0147\right)$, when $\delta_T = \pm\frac{1}{2}$.

A class of algorithms is suggested from optimally combining the two estimators $\widehat{\delta}_{1T}$ and $\widehat{\delta}_{-1T}$. One such estimator replaces step 3 above with

$$\widehat{\delta}_T = \frac{\widehat{\delta}_{1T} + \widehat{\delta}_{-1T}}{2} + g(\widehat{\delta}_{1T}^2) - g(\widehat{\delta}_{-1T}^2),$$

where

$$g(x) = \frac{1}{4}\log(3x^2 + 6x + 1) - \frac{\sqrt{6}}{24}\log\left(\frac{x + 1 - \sqrt{\frac{2}{3}}}{x + 1 + \sqrt{\frac{2}{3}}}\right).$$

An asymptotically equivalent estimator is obtained by using, with $\overline{\delta}_T$ equal to the FTI1 estimator,

$$\widehat{\delta}_T = \frac{\widehat{\delta}_{1T} + \widehat{\delta}_{-1T}}{2} + \left(\widehat{\delta}_{1T} - \widehat{\delta}_{-1T}\right)\frac{3\overline{\delta}_T^3 + 2\overline{\delta}_T}{3\overline{\delta}_T^4 + 6\overline{\delta}_T^2 + 1}.$$

The new estimator $\widehat{\omega}_T$ satisfies the same type of central limit theorem, but with

$$v_T^2 = \frac{8\pi^3 f(\omega)}{\rho^2}\frac{\pi^2\delta_T^2}{\sin^2(\pi\delta_T)}\frac{\left(1 - \delta_T^2\right)^2\left(3\delta_T^4 + 1\right)}{3\delta_T^4 + 6\delta_T^2 + 1}.$$

The new asymptotic efficiency relative to the periodogram maximizer is largest $\left(\pi^2/6 \sim 1.6449\right)$ when $\delta_T = 0$ and smallest $\left(57\pi^4/5504 \sim 1.0088\right)$ when $\delta_T = \pm\frac{1}{2}$. For further details, see Quinn and Hannan (2001).

Quinn (2006) has developed and analyzed algorithms which may be applied to time series that have been "tapered."

11. Estimation using only the moduli of the DFT

Popular estimation techniques have been based on fitting curves to the periodogram. The most well-known estimator is the quadratic interpolator, which is defined as the maximizer of the quadratic fitted through the points $\left\{\left(\omega_{k_T+j}, I_X(\omega_{k_T+j})\right); j = -1, 0, 1\right\}$. It is shown in Q&H that the frequency estimator is given by

$$\widehat{\omega}_T = 2\pi\frac{k_T + \widehat{\delta}_T}{T},$$

where

$$\begin{aligned}\widehat{\delta}_T &= \frac{1}{2}\frac{I_X\left(\omega_{k_T+1}\right) - I_X\left(\omega_{k_T-1}\right)}{2I_X\left(\omega_{k_T}\right) - I_X\left(\omega_{k_T-1}\right) - I_X\left(\omega_{k_T+1}\right)} \\ &= \delta_T + \frac{4\delta_T^3 - \delta_T}{1 - 3\delta_T^2} + o(1)\end{aligned}$$

almost surely as $T \to \infty$, when $\delta_T \in \left[aT^{-\nu} - \frac{1}{2}, \frac{1}{2} - aT^{-\nu}\right]$, for fixed $a > 0$ and $0 < \nu < 1/2$, instead of δ_T. Therefore, $T(\widehat{\omega}_T - \omega)$ does not converge almost surely to 0, and the quadratic interpolator has an unacceptably large bias. Other estimators have been suggested, such as (Hawkes, 1990)

$$\widehat{\delta}_T = c\frac{\sqrt{I_X\left(\omega_{k_T+1}\right)} - \sqrt{I_X\left(\omega_{k_T-1}\right)}}{\sqrt{I_X\left(\omega_{k_T}\right)} + \sqrt{I_X\left(\omega_{k_T-1}\right)} + \sqrt{I_X\left(\omega_{k_T+1}\right)}}, \tag{20}$$

where c does not depend on δ. All such estimators, however, will also have the same bias problems. To see this, let

$$\begin{aligned}\widehat{\delta}_T &= \frac{a\sqrt{I_X\left(\omega_{k_T+1}\right)} + b\sqrt{I_X\left(\omega_{k_T-1}\right)}}{c\sqrt{I_X\left(\omega_{k_T}\right)} + d\sqrt{I_X\left(\omega_{k_T-1}\right)} + e\sqrt{I_X\left(\omega_{k_T+1}\right)}} \\ &= \frac{a\left|\frac{\delta_T}{\delta_T-1}\right| + b\left|\frac{\delta_T}{\delta_T+1}\right|}{c + d\left|\frac{\delta_T}{\delta_T-1}\right| + e\left|\frac{\delta_T}{\delta_T-1}\right|} + o(1),\end{aligned}$$

almost surely as $T \to \infty$, again when $\delta_T \in \left[aT^{-\nu} - \frac{1}{2}, \frac{1}{2} - aT^{-\nu}\right]$, for fixed $a > 0$ and $0 < \nu < 1/2$. In order that $\widehat{\delta}_T - \delta_T$ converge almost surely to 0, it is necessary that in some neighborhood of 0,

$$\delta = \frac{a\left|\frac{\delta}{\delta-1}\right| + b\left|\frac{\delta}{\delta+1}\right|}{c + d\left|\frac{\delta}{\delta-1}\right| + e\left|\frac{\delta}{\delta+1}\right|}.$$

Thus, for $\delta > 0$,

$$\begin{aligned}\delta &= \frac{a\left(\frac{\delta}{1-\delta}\right) + b\left(\frac{\delta}{\delta+1}\right)}{c + d\left(\frac{\delta}{1-\delta}\right) + e\left(\frac{\delta}{\delta+1}\right)} \\ &= \frac{a\delta(1+\delta) + b\delta(1-\delta)}{c\left(1-\delta^2\right) + d\delta(1+\delta) + e\delta(\delta+1)},\end{aligned}$$

so that

$$a = d, b = -e, c = d - e,$$

while, for $\delta < 0$,

$$\begin{aligned}\delta &= \frac{a\left(\frac{-\delta}{1-\delta}\right) + b\left(\frac{-\delta}{\delta+1}\right)}{c + d\left(\frac{-\delta}{1-\delta}\right) + e\left(\frac{-\delta}{\delta+1}\right)} \\ &= \frac{-a\delta(1+\delta) - b\delta(1-\delta)}{c\left(1-\delta^2\right) - d\delta(1+\delta) - e\delta(\delta+1)},\end{aligned}$$

which implies that

$$a = d, b = -e, c = e - d.$$

The two sets of conditions can hold only when $e = d$, and therefore, $c = 0$ and $b = -a$. Thus, the only estimator with the correct order of consistency is formed using

$$\widehat{\delta}_T = \frac{\sqrt{I_X\left(\omega_{k_T+1}\right)} - \sqrt{I_X\left(\omega_{k_T-1}\right)}}{\sqrt{I_X\left(\omega_{k_T-1}\right)} + \sqrt{I_X\left(\omega_{k_T+1}\right)}}.$$

This estimator has, however, very poor asymptotic properties. The main problem with estimators that use only the *moduli* of the Fourier coefficients, and not the arguments, or phases, is that they are missing important sign information.

Rife and Vincent (1970) suggested the estimator

$$\widehat{\omega}_T = 2\pi \frac{k_T + \widehat{\delta}_T}{T}$$

$$\widehat{\delta}_T = \widehat{\alpha}_T \frac{\sqrt{I_X\left(\omega_{k_T+\widehat{\alpha}_T}\right)}}{\sqrt{I_X\left(\omega_{k_T}\right)} + \sqrt{I_X\left(\omega_{k_T+\widehat{\alpha}_T}\right)}}$$

$$\widehat{\alpha}_T = \operatorname{sgn}\left\{I_X\left(\omega_{k_T+1}\right) - I_X\left(\omega_{k_T-1}\right)\right\}.$$

The motivation behind this is that δ_T is more likely to be positive if $I_X(\omega_{k_T+1}) > I_X(\omega_{k_T-1})$ and vice versa. The resulting frequency estimator exhibits bizarre behavior. It is shown in Q&H that if $\omega/(2\pi)$ is irrational, $T^{5/4}(\widehat{\omega}_T - \omega)$ does not converge in probability to 0, whereas if $\omega/(2\pi)$ is rational, $T^{3/2}(\widehat{\omega}_T - \omega)$ converges in distribution, but is not asymptotically normal. The problems with the estimator result from making the wrong choice in $\widehat{\alpha}_T$ when δ_T is close to 0. This can be partially corrected by using

$$\operatorname{sgn} \operatorname{Re} \frac{Y_{k_T+j}}{Y_{k_T}}, \quad j = -1, 1.$$

The details are given in Q&H. It is conjectured that no estimation procedure based *only* on k_T and $\left\{I_X(\omega_{k_T+j}) j = -1, 0, 1\right\}$ will have the same order of consistency as those which use the additional information provided by the Fourier coefficients.

12. More than one sinusoid

The model containing several sinusoids

$$X_t = \mu + \sum_{j=1}^{f} \rho_j \cos\left(\omega_j t + \phi_j\right) + \varepsilon_t \tag{21}$$

is reasonable whenever data containing different sinusoids with different and unrelated frequencies are added, or when a time series is thought to have been produced by some periodic, but not necessarily sinusoidal phenomenon. In the latter case, the frequencies ω_j above will be harmonically related, that is, integer multiples of a fundamental frequency. In practice, and especially in the sonar context, a time series may be the noisy sum of a large number of sinusoids, with some frequencies harmonically related and others not related at all. In this section, we consider f to be known.

The least squares estimators of the ω_j are obtained (Bloomfield, 1976; Quinn and Hannan, 2001) by minimizing with respect to the ω_j

$$\min_{\mu,\rho_1,\ldots,\rho_f,\phi_1,\ldots,\phi_f} \sum_{t=0}^{T-1} \left\{ X_t - \mu - \sum_{j=1}^{f} \rho_j \cos\left(\omega_j t + \phi_j\right) \right\}^2,$$

or, equivalently,

$$\min_{\mu,\alpha_1,\ldots,\alpha_f,\beta_1,\ldots,\beta_f} \sum_{t=0}^{T-1} \left[X_t - \mu - \sum_{j=1}^{f} \left\{ \alpha_j \cos\left(\omega_j t\right) + \beta_j \sin\left(\omega_j t\right) \right\} \right]^2. \tag{22}$$

The latter function is easily computed by regression, for fixed $\omega_1, \ldots, \omega_f$, as in the single frequency case, and is asymptotically equivalent to

$$\sum_{t=0}^{T-1} \left(X_t - \overline{X}\right)^2 - \sum_{j=1}^{f} I_X(\omega_j). \tag{23}$$

Consequently, the least squares estimators are, at least in the usual sense, asymptotically equivalent to local maximizers of the periodogram. It has long been the approach to estimate several frequencies by looking for the largest local maxima of the periodogram. However, there are several problems with this approach.

1. Two sinusoids might have frequencies "close together" relative to T, invalidating the approximation of (22) by (23).
2. The sidelobes from one sinusoid might be interpreted as coming from separate sinusoids if one sinusoid has a much larger amplitude than some others.

The two problems are quite different and have quite different solutions. It is the second problem that has driven the development of MUSIC and similar techniques – the fact that the periodogram cannot "resolve" several frequencies if the amplitudes of the sinusoids are quite different. However, there is no inherent reason why the frequencies of several sinusoids should be resolved by maximizing a function of a *single* frequency.

12.1. The resolution of close frequencies

Hannan and Quinn (1989) consider the case where $f = 2$ and $\omega_2 = \omega_1 + T^{-1}a$, where ω_1 and a are fixed. The T^{-1} term might seem odd, but something is needed to model

the case of "close" frequencies, and T^{-1} is suggested by the analysis. The case where $\omega_1 = 0$ is considered special, as it can be argued that then there are three sinusoids with close frequencies at 0 and $\pm T^{-1}a$. Equation (22) may be shown in the case where $\omega_1 = \omega \neq 0$ and $\omega_2 = \omega + T^{-1}a$ to equal

$$\sum_{t=0}^{T-1} \left(X_t - \overline{X}\right)^2 - \frac{1}{1 - \frac{\sin^2(a/2)}{(a/2)^2}} \left[I_X(\omega) + I_X(\omega + T^{-1}a) - \frac{2}{T} \operatorname{Re} \left\{ \overline{Y}_T(\omega + T^{-1}a) Y_T(\omega) \frac{e^{ia} - 1}{ia} \right\} \right],$$

where

$$Y_T(\omega) = \sum_{t=0}^{T-1} X_t e^{-i\omega t}.$$

The regression sum of squares, as a function of ω and a, is thus easily computed using the discrete Fourier transform. Hannan and Quinn (1989) and Quinn and Hannan (2001) show that the least squares estimators $\widehat{\omega}_T$ and $\widehat{a}_T$ are such that

$$\begin{aligned} T\left(\widehat{\omega}_T - \omega\right) &\to 0 \\ \widehat{a}_T - a &\to 0, \end{aligned}$$

almost surely as $T \to \infty$, and that $[T^{3/2}\left(\widehat{\omega}_T - \omega\right)\ T^{1/2}\left(\widehat{a}_T - a\right)]'$ is asymptotically normally distributed with zero mean and a covariance matrix, which depends in a complicated way on $\phi_2 - \phi_1 - a/2, a/2, \rho_1, \rho_2$, and ω. They also discuss the related problem of estimating a single low frequency of the form a/T. Consider the case

$$f = 2,\ \rho_1 = 1,\ \rho_2 = 0.5,\ \phi_1 = \phi_2 = 0,\ \omega_1 = \frac{2\pi 135.3}{1024},\ a = 0.9,\ T = 1024,$$

where $\{\varepsilon_t\}$ is Gaussian and white, with common variance 1. The periodogram, shown in Fig. 7, does not resolve the two frequencies, as the first right sidelobe due to the sinusoid at ω_1 has been confused with the main lobe of the sinusoid at $\omega_1 + T^{-1}a$.

The MUSIC spectrum is, with $K = 100$, shown in Fig. 8.

MUSIC does not resolve the two close frequencies. However, if we let $\zeta\left(\omega_1, \omega_2\right)$ be the regression sum of squares, computed for example by

$$\zeta\left(\omega_1, \omega_2\right) = \frac{1}{1 - \frac{\sin^2(a/2)}{(a/2)^2}} \left[I_X(\omega_1) + I_X(\omega_2) - \frac{2}{T} \operatorname{Re} \left\{ \overline{Y}_T(\omega_2) Y_T(\omega_1) \frac{e^{ia} - 1}{ia} \right\} \right],$$

where $a = T(\omega_2 - \omega_1)$, the frequencies may be resolved (Fig. 9). Although it is difficult to see this in a single surface plot of ζ, the plot of

$$\xi\left(\omega\right) = \max_{\omega_1} \zeta\left(\omega_1, \omega\right)$$

is convincing (Fig. 10).

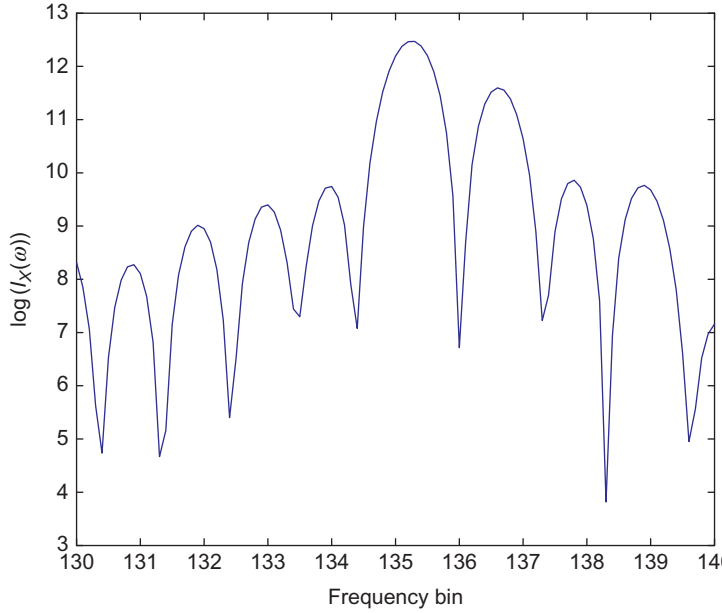

Fig. 7. Periodogram, two close frequencies.

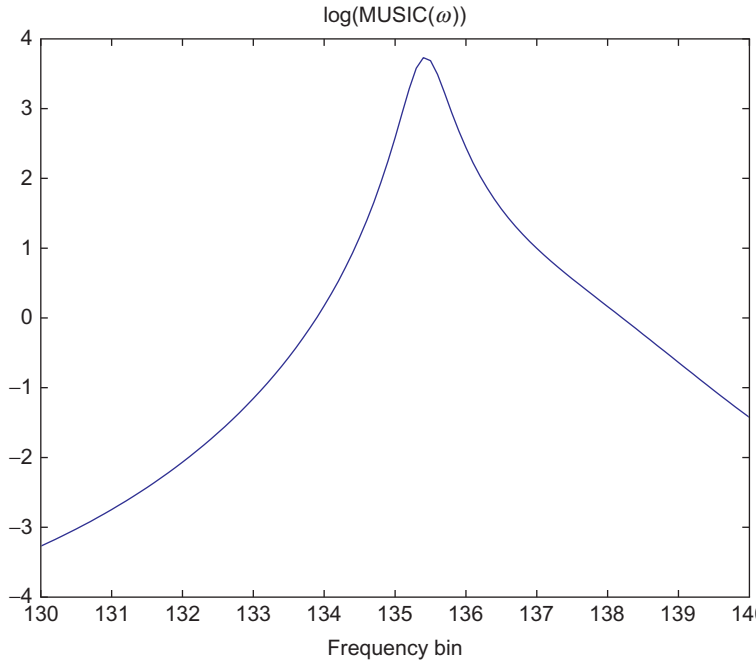

Fig. 8. MUSIC, two close frequencies.

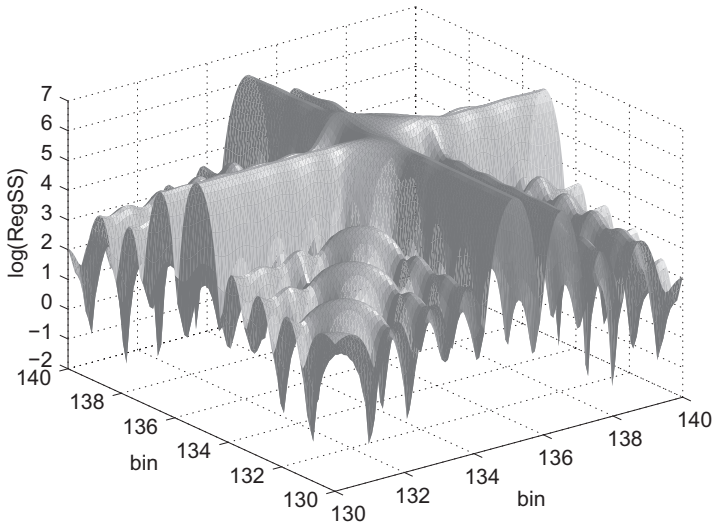

Fig. 9. Log of regression sum of squares for two close frequencies.

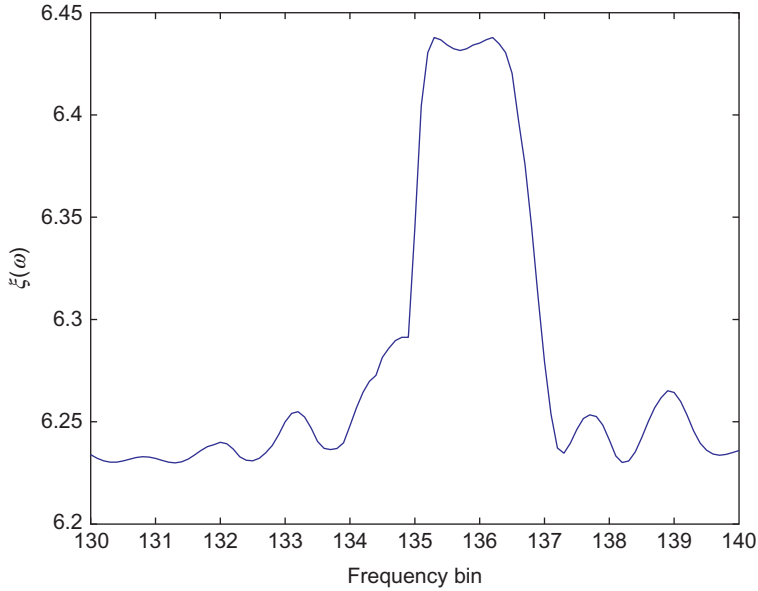

Fig. 10. Resolution using ξ.

A discussion of resolution is not complete without the mention of "sidelobe suppression" techniques. Figure 11 is a plot of the periodogram after applying a Hann window. This has, however, resulted in quite a bit of bias.

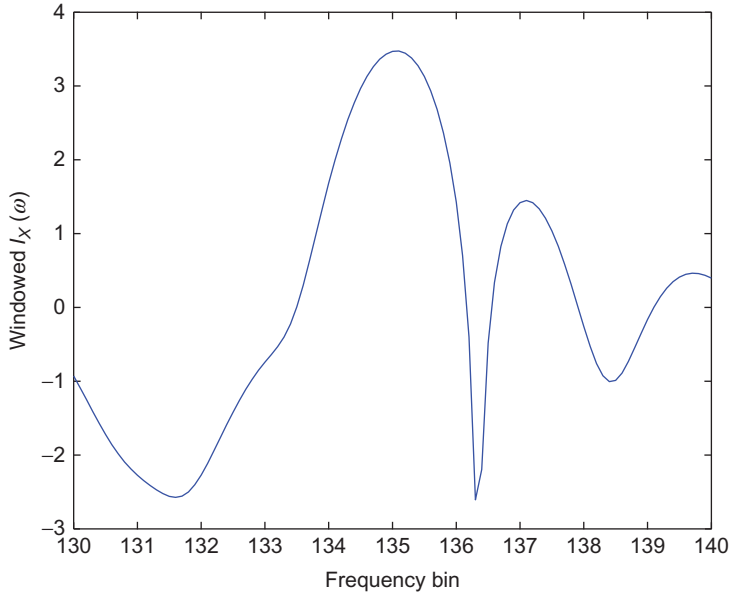

Fig. 11. Sidelobe suppression using Hanning window.

12.2. *Other resolution problems*

Consider the case where $f = 2, \rho_1 = 10, \rho_2 = 1, \omega_1 = 171\pi/1024, \omega_2 = 190\pi/1024$, $\phi_1 = \phi_2 = 0$, and $T = 1024$. We shall assume first that $\{\varepsilon_t\}$ is white, with common variance 1. The log of the periodogram of a simulated time series is partially shown in Fig. 12. The bin number k refers to the frequency $2\pi k/T$. It is clearly seen that the second frequency corresponds to the sixth largest local maximum.

In contrast, the relevant section of the MUSIC spectrum is given in Fig. 13, with $K = 100$. Clearly, MUSIC does not resolve the two frequencies, even though they are separated.

The situation may even be worse if the noise is not white. Suppose, now that $f = 2$, $\rho_1 = 1, \rho_2 = 1, \omega_1 = 171\pi/1024, \omega_2 = 190\pi/1024, \phi_1 = \phi_2 = 0$, and $T = 1024$, so that the amplitudes are equal, but that the noise is an autoregression of order 2, satisfying

$$\varepsilon_t - 1.64\varepsilon_{t-1} + 0.81\varepsilon_{t-2} = u_t,$$

where $\{u_t\}$ is white, with common variance 1. Thus $\{\varepsilon_t\}$ exhibits a pseudo-cycle near the frequency $140\pi/1024$. The local periodogram of a simulated time series is shown in Fig. 14.

Not only the peak in the periodogram at ω_2 has been suppressed, there are many other spurious peaks because the background spectral density is not flat. In fact, the periodogram at ω_1 is only marginally larger than the periodogram near frequency bin 73.6. This is because of the thresholding effect: the largest periodogram value "due to noise" is larger than the periodogram at the true frequency. A discussion is given in Quinn and Kootsookos (1994) of the complex Gaussian white noise case.

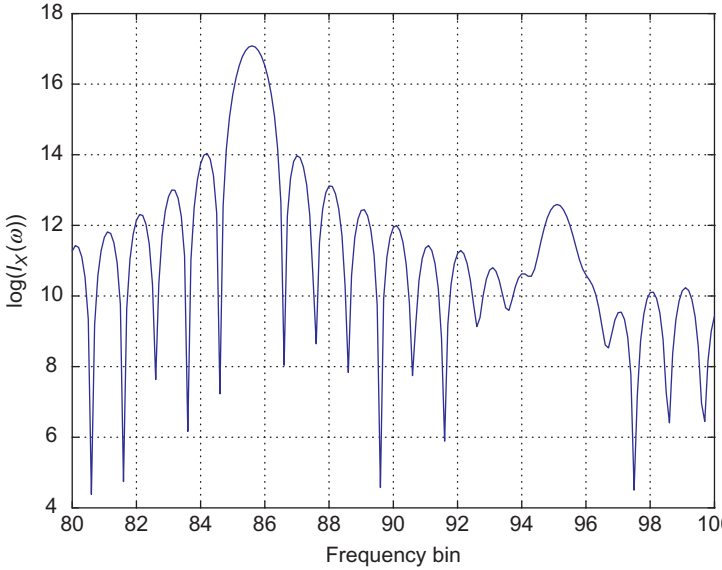

Fig. 12. Disparate amplitudes.

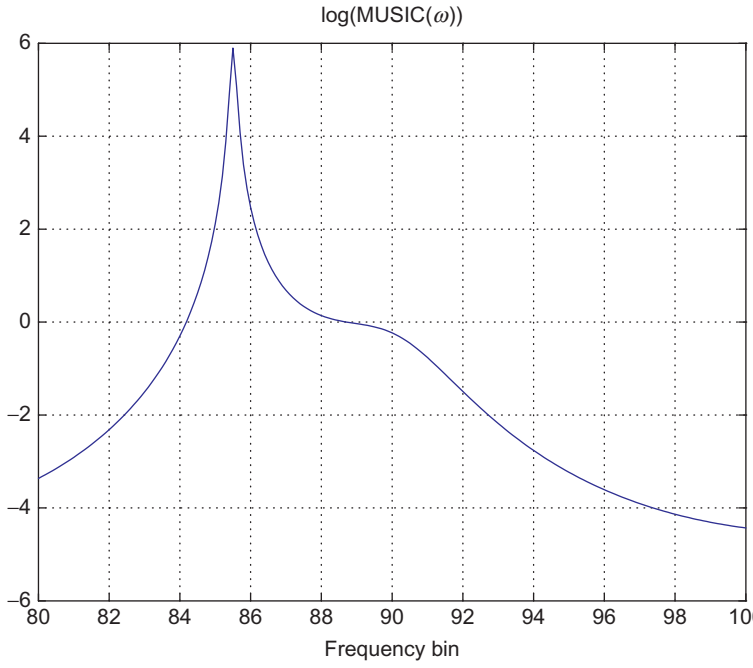

Fig. 13. MUSIC, disparate amplitudes.

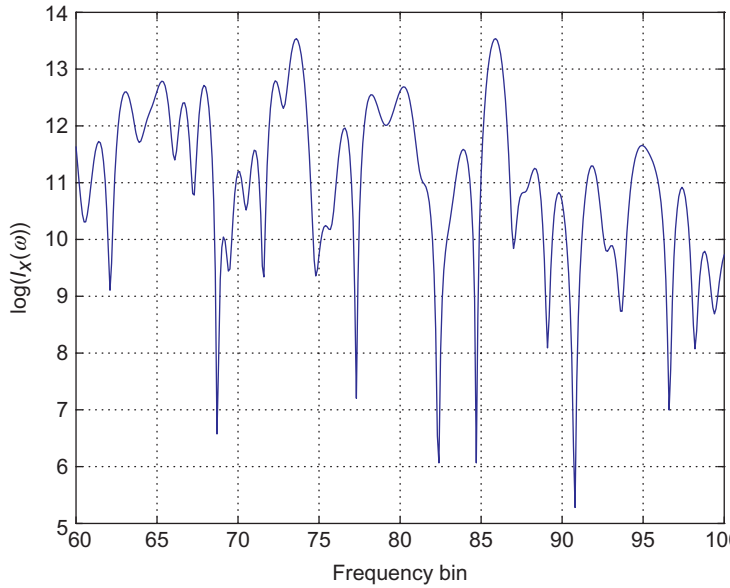

Fig. 14. Periodogram, same amplitudes and colored noise.

Bloomfield (1976) and Quinn and Hannan (2001) suggest that, rather than trying to resolve frequencies by using the periodogram, sinusoidal terms be removed by regression after being detected and then estimated, one by one. This approach will work as long as the frequencies are separated, but not when they are close. To reduce problems that arise as a result of non-whiteness of the background noise, Quinn (2004) has developed a technique that simultaneously estimates an autoregressive approximation to the noise and "equalises" its effects, so that the periodogram appears to be from the sum of sinusoids and white noise. An extension to the complex case is given by Quinn (2007).

12.3. Harmonically related frequencies

When the sinusoidal sum is actually the approximation to a periodic function, the frequencies in (21) are such that $\omega_j = j\omega$, where ω is the unknown "fundamental" frequency. The regression sum of squares based on white noise assumptions is then not the appropriate function to maximize, as the component sinusoids are influenced individually by the noise spectral densities at the harmonics, and each sinusoid is a function of ω. If $\{\varepsilon_t\}$ is Gaussian, with known spectral density, then an asymptotically equivalent approach to maximum likelihood is (Quinn and Thomson, 1991) to maximize

$$\sum_{j=1}^{f} \frac{I_X(j\omega)}{f(j\omega)}. \tag{24}$$

The spectral density, of course, is not known and must be estimated. Quinn and Thomson suggest estimating $f(\lambda)$ by using the median of the $I_X(\omega)$ near λ.

Chiu (1989) uses trimmed means. Another approach is to model the noise: Quinn and Thomson (1998) have assumed $\{\varepsilon_t\}$ to be autoregressive and estimate the sinusoidal and autoregressive parameters simultaneously. Their algorithm also allows missing values in the data by incorporating an EM-like algorithm. In all cases, $\widehat{\omega}$, the maximizer of (24), is such that $T^{3/2}(\widehat{\omega}-\omega)$ is asymptotically normally distributed with zero mean and variance

$$\frac{48\pi}{\sum_{j=1}^{f}}\frac{j^2\rho_j^2}{f(j\omega)}.$$

Semiparametric or nonparametric approaches to the problem of estimating the period of a periodic function have recently been proposed by Gassiat and Lévy-Leduc (2006), Hall et al. (2000), Hall and Li (2006), Hall and Yin (2003),, and Lévy-Leduc et al. (2008). These approaches apply to the more general case where the time series is sampled irregularly.

13. Complex sinusoids

Many engineering techniques have been explicitly developed for complex-valued processes satisfying equations of the form

$$X_t = \mu + \sum_{j=1}^{f} D_j \exp\left(i\omega_j t\right) + \varepsilon_t, \tag{25}$$

where the D_j are complex and $\{\varepsilon_t\}$ a complex-valued stationary process. Many of the techniques mentioned above are easily modified, if they need to be. There is a class of techniques, however, which is only applicable to complex sinusoids – those which use only the arguments (phases) of the time series and ignore the moduli. The need for such techniques arises in systems where the moduli may be distorted, for example, by Automatic Gain Control, but the arguments are distortion free.

Consider $\{X_t\}$ satisfying (25) with $\mu = 0$ and $f = 1$

$$\begin{aligned} X_t &= D\exp(i\omega t) + \varepsilon_t \\ &= D\exp(i\omega t)\left\{1 + D^{-1}\exp(-i\omega t)\,\varepsilon_t\right\}. \end{aligned}$$

If $\{\varepsilon_t\}$ is complex Gaussian and white, then so is $\{v_t\}$, where $v_t = 1 + D^{-1}\exp(-i\omega t)\,\varepsilon_t$. Thus,

$$\arg X_t = \arg D + \omega t + \arg v_t \mod(2\pi). \tag{26}$$

Several popular techniques are based on the fact that this equation is "linear" in t, and thus, linear regression techniques may be used. The fact that "wrapping" occurs

with increasing t has led several authors to take first differences, or args of ratios, obtaining

$$\arg\left(\frac{X_t}{X_{t-1}}\right) = \omega + \arg\left(\frac{v_t}{v_{t-1}}\right) \mod(2\pi)$$

and then estimating ω by weighted averages of the $\arg\left(\frac{X_t}{X_{t-1}}\right)$. Such techniques have been shown to be inconsistent (Quinn, 2000). Although it is tempting to believe that a simple technique exists which uses (26) and is both consistent and asymptotically close to efficient, the wrapping problem seems insurmountable. Recently, McKilliam et al. (2010) have analyzed the estimator of ω found by minimizing

$$\sum_{t=0}^{T-1}\left\langle\frac{\arg X_t - \arg D - \omega t}{2\pi}\right\rangle^2$$

with respect to $\arg D$ and ω, where $\langle x\rangle = x - \lfloor x\rceil$ and so $|\langle x\rangle|$ is the distance between x and its nearest integer. Although the asymptotic behavior is excellent, the technique is computationally intensive.

14. Related problems and areas

The problems of testing for the presence of sinusoids, and estimating the number of sinusoids, have not been addressed. The reader is referred to Quinn and Hannan (2001). The state of the art in estimating the number of sinusoids is most likely Kavalieris and Hannan (1994), who use a BIC-like procedure to estimate both the number of sinusoids and the order of the best autoregressive fit to the noise. The problem of "tracking" an evolving frequency is also not discussed, although there is an enormous literature on this topic. A related problem is the estimation and tracking of the direction of arrival of a signal, using arrays of sensors.

References

Bartlett, M.S., 1967. Inference and stochastic processes. J. Roy. Statist. Soc. A 130, 457–477.

Bloomfield, P., 1976. Fourier Analysis of Time Series: An Introduction, second ed., 2000. Wiley, New York.

Brillinger, D.R., 1974. Time Series Data Analysis and Theory, expanded ed., 1981. Holden-Day, San Francisco.

Brillinger, D.R., 1987. Fitting cosines: some procedures and some physical examples. In: MacNeill, I.B. Umphrey, G.J. (Eds.), Applied Probability, Stochastic Processes, and Sampling Theory. Reidel, Dordrecht, pp. 75–100..

Buys-Ballot, C., 1847. Les Changements Périodiques de Temperature. Kemink et Fils, Utrecht.

Chen, Z.G., Wu, K.H., Dahlhaus, R., 2000. Hidden frequency estimation with data tapers. J.Time Series Anal. 21, 113–142.

Chiu, S.-T., 1989. Detecting periodic components in a white Gaussian time series. J. Roy. Statist. Soc. B 51, 249–259.

Cooley, J.W., Tukey, J.W., 1965. An algorithm for the machine computation of complex Fourier series. Math. Comput. 19, 297–301.

de Prony, G.R., 1795. Essai éxperimental et analytique: sur les lois de la dilatabilité de fluides élastique et sur celles de la force expansive de la vapeur de l'alkool, à différentes températures. Journal de l'École Polytechnique, 1, cahier 22, 24–76.

Fernandes, J.M., Goodwin, G.C., De Souza, C.E., 1987. Estimation of models for systems having deterministic and random disturbances. Proc. 10th World Cong. Automat. Contr. 10, 370–375.

Gassiat, E., Lévy-Leduc, C., 2006. Efficient semiparametric estimation of the periods in a superposition of periodic functions with unknown shape. J. Time Series Anal. 27, 877–910.

Hall, P., Li, M., 2006. Using the periodogram to estimate period in nonparametric regression. Biometrika 93, 411–424.

Hall, P., Reimann, J., Rice, J., 2000. Nonparametric estimation of a periodic function. Biometrika 87, 545–557.

Hall, P., Yin, J., 2003. Nonparametric methods for deconvolving multiperiodic functions. J. Roy. Statist. Soc. B 65, 869–886.

Hannan, E.J., 1973. The estimation of frequency. J. Appl. Prob. 10, 510–519.

Hannan, E.J., Quinn, B.G., 1989. The resolution of closely adjacent spectral lines. J. Time Series Anal. 10, 13–22.

Hawkes, K., 1990. Bin interpolation. Technically Speaking, ESL Inc., pp. 17–30.

Heideman, M.T., Johnson, D.H., Burrus, C.S., 1984. Gauss and the history of the fast Fourier transform. IEEE ASSP magazine 1, 14–21.

Johnson, D.H., 1982. The application of spectral estimation methods to bearing estimation problems. Proc. IEEE 70, 1018–1028.

Kavalieris, L., Hannan, E.J., 1994. Determining the number of terms in a trigonometric regression., J. Time Series Anal. 15, 613–625.

Lévy-Leduc, C., Moulines, E., Roueff, F., 2008. Frequency estimation based on the cumulated Lomb-Scargle periodogram. J. Time Series Anal. 29, 1104–1131.

Li, T.H., Kedem, B., 1993. Strong consistency of the contraction mapping method for frequency estimation. IEEE Trans. Inf. Theor. 39, 989–998.

MacLeod, M.D., 1998. Fast nearly ML estimation of the parameters of real or complex single tones or resolved multiple tones. IEEE Trans. Signal Process. 46, 141–148.

McKilliam, R.G., Quinn, B.G., Clarkson, I.V.L., Moran, B., 2010. Frequency estimation by phase unwrapping. IEEE Trans. Signal Process. 58, 2953–2963.

Nehorai, A., Porat, B., 1986. Adaptive comb filtering for harmonic signal enhancement, IEEE Trans. ASSP. 34, 1124–1138.

Pisarenko, V.F., 1973. The retrieval of harmonics from a covariance function. Geophys. J.R. Astr. Soc. 10, 347–366.

Priestley, M.B., 1981. Spectral Analysis and Time Series. Academic Press, London.

Quinn, B.G., 1992. Some new high-accuracy frequency estimators. Proceedings of ISSPA, Gold Coast, pp. 323–326.

Quinn, B.G., 1994. Estimating frequency by interpolation using Fourier coefficients. IEEE Trans. Signal Process. 42, 1264–1268.

Quinn, B.G., 2000. On Kay's frequency estimator. J. Time Series Anal. 21, 707–712.

Quinn, B.G., 2004. Estimating a sinusoid in low SNR coloured noise. Proceedings of the 2004 Intelligent Sensors, Sensor Networks & Information Processing Conference, IEEE Press, Melbourne, Australia, pp. 301–306.

Quinn, B.G., 2006. Frequency estimation using tapered data. Proceedings of the 2006 International Conference on Acoustics, Speech and Signal Processing, III, IEEE Press, New York, pp. 73–76.

Quinn, B.G., 2007. Efficient estimation of the parameters in a sum of complex sinusoids in complex autoregressive noise. Proceedings of the 2007 Asilomar conference on Signals, Systems and Computers, IEEE Press, New York, pp. 636–640.

Quinn, B.G., Fernandes, J.M., 1991. A fast efficient technique for the estimation of frequency. Biometrika 78, 489–498.

Quinn, B.G., Hannan, E.J., 2001. The Estimation and Tracking of Frequency. Cambridge University Press, New York.

Quinn, B.G., Kootsookos, P.J., 1994. Threshold behavior of the maximum likelihood estimator of frequency. IEEE Trans. Signal Process. 42, 3291–3294.

Quinn, B.G., McKilliam, R.G., Clarkson, I.V.L., 2008. Maximizing the periodogram. Proc. IEEE Globecom 2008, 1–4.

Quinn, B.G., Thomson, P.J., 1991. Estimating the frequency of a periodic function. Biometrika 78, 65–75.

Quinn, B.G., Thomson, P.J., 1998. Fitting mixed sinusoidal/AR models to time series with missing values, Unpublished manuscript, Seminar delivered at the University of Kent.

Rice, J.A., Rosenblatt, M., 1988. On frequency estimation. Biometrika 74, 477–484.

Rife, D.C., Boorstyn, R.R., 1974. Single tone parameter estimation from discrete-time observations. IEEE Trans. Inf. Theor. 20, 591–598.

Rife, D.C., Vincent, G.A., 1970. Use of the discrete Fourier transform in the measurement of frequencies and levels of tones. Bell Syst. Tech. J. 49, 197–228.

Sakai, H., 1984. Statistical analysis of Pisarenko's method for sinusoidal frequency estimation. IEEE Trans. ASSP. 32, 95–101.

Schmidt, R.O., 1981. A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation. Ph.D. thesis, Stanford University, CA.

Schmidt, R.O., 1986. Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. 34, 276–280.

Schuster, A., 1898. On the investigation of hidden periodicities with application to the supposed 26-day period of meteorological phenomena. Terr. Magn. Atmos. Electr. 3, 13–41.

Song, K.-S., Li, T.-H., 2000. A statistically and computationally efficient method for frequency estimation. Stochastic Processes and Their Applications 86, 29–47.

Truong-Van, B., 1990. A new approach to frequency analysis with amplified harmonics. J. Roy. Statist. Soc. B 52, 203–222.

Walker, A.M., 1971. On the estimation of the harmonic component in a time series with stationary independent residuals. Biometrika 58, 21–36.

Whittaker, E., Robinson, G., 1944. The Calculus of Observations, fourth ed. Blackie, London.

Whittle, P., 1952. The simultaneous estimation of a time series harmonic components and covariance structure. Trabajos de Estadistica 3, 43–57.

Yule, G.U., 1927. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's Sunspot Numbers. Phil. Trans. Roy. Soc. London A 226, 267–298.

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00022-3

A Wavelet Variance Primer

Donald B. Percival [1] *and Debashis Mondal* [2]

[1]*Applied Physics Laboratory, University of Washington, Seattle, WA, USA*
[2]*Department of Statistics, University of Chicago, South University Avenue, Chicago, IL, USA*

Abstract

The wavelet variance is a decomposition of the variance of a time series. Because of its scale-based nature, the wavelet variance offers insight into various time series, particularly in the physical sciences. This primer is a basic introduction to the wavelet variance, starting with its definition in terms of the discrete wavelet transform, proceeding with a discussion of the large-sample statistical properties of its basic estimators, and then continuing with an examination of estimators appropriate for time series with either missing values or contamination by discordant values. The discussion then moves to two uses of the wavelet variance involving its across-scale patterns, namely, estimation of exponents of power-law processes and estimations of characteristic scales. The primer closes with examples of the wavelet variance applied to time series involving atomic clocks, sea-ice thickness, the albedo of Arctic ice, X-ray fluctuations from binary stars, and coherent structures in river flow.

Keywords: analysis of variance, characteristic scales, discrete wavelet transform, Daubechies wavelet filters, intrinsically stationary time series, multiscale contamination, power-law processes, missing observations, robust estimator.

1. Introduction

The discrete wavelet transform (DWT), as formulated in the late 1980s by Daubechies (1988), Mallat (1989a,b,c), and others, has inspired extensive research into how to use this transform to study time series. One focus of this research has been on the wavelet variance (also called the wavelet spectrum). The wavelet variance decomposes the variance of a time series and, hence, provides an analysis of variance (ANOVA).

The most widely used ANOVA technique in time series analysis is spectral analysis, which involves the Fourier-based spectral density function (SDF). The ANOVA that the wavelet variance provides is in many ways similar to that afforded by the SDF (Li and Oh, 2002); however, from a practitioner's point of view, there are key differences. The SDF is a decomposition of variance across a continuum of Fourier frequencies. Each component in the decomposition reflects the degree to which a time series resembles a sinusoid with a particular frequency. The wavelet variance differs in that it is a decomposition across a discrete set of scales. Roughly speaking, a scale is an interval (or span) of time over which a time series is averaged. The strength of each component in the decomposition measures how much variability there is between adjacently located averages associated with a particular scale. The concept of scale is distinct from that of period (the inverse of frequency). Both are measured in the same units, but period does not involve averaging. Although it is possible to estimate an SDF indirectly via the wavelet variance (Tsakiroglou and Walden, 2002), the different interpretations that frequency and scale have make the ANOVA afforded by the wavelet variance more appealing than the one given by the SDF for interpreting certain time series. Examples of applications that have made use of the wavelet variance are extensive and include time series related to electroencephalographic sleep state patterns of infants (Chiann and Morettin, 1998), frequency instability of atomic clocks (Greenhall et al., 1999), rainfall/runoff relationships (Labat et al., 2001), variations in soil composition (Lark and Webster, 2001), ocean surface waves (Massel, 2001), surface albedo and temperature in desert grasslands (Pelgrum et al., 2000), heart rate variability (Pichot et al., 1999), stochastic fluctuations on accreting binary stars (Scargle et al., 1993), solar coronal activity (Rybák and Dorotovič, 2002), and the El Niño–Southern Oscillation (Torrence and Compo, 1998). In addition, in contrast to the Fourier transform, the DWT is localized in time, and hence, the wavelet variance can be readily adapted for exploring processes that are locally stationary with time-varying SDFs (Nason et al., 2000) and for detecting inhomogeneities in time series (Whitcher et al., 2002).

The intent of this chapter is to provide a basic introduction to the wavelet variance, with an emphasis on its interpretation, its statistical properties, and some recent extensions to the basic methodology. We start with an overview of the maximal overlap DWT (MODWT), which is the version of the DWT of most interest for formulating the wavelet variance (Section 2). The MODWT leads to a basic ANOVA of a time series, which we describe in Section 3. If we assume that the time series under analysis is a realization of an intrinsically stationary process (as defined in Section 4), we can define a theoretical wavelet variance and regard the descriptive statistics discussed in Section 3 as basic estimators of this variance. We discuss the fundamental statistical theory behind estimators of the wavelet variance in Section 5, following which we discuss estimators intended to handle special circumstances (gappy time series in Section 6.1 and series with aberrant observations in Section 6.2). We then describe two wavelet-based methodologies in Section 7, one for deducing the presence of power-law dependence in a time series and the other for defining a characteristic scale. In both cases, the statistics that arise are qualitatively similar in that they combine wavelet variance estimates together across adjacent scales. We devote the penultimate Section (8) to five real-world examples illustrating the methodology discussed in previous sections (two of these examples serve to briefly compare Fourier-based spectral analysis with

the analysis afforded by the wavelet variance; see also Faÿ et al. (2009)). We close with some concluding remarks in Section 9.

2. Maximal overlap discrete wavelet transform

The wavelet variance is based on the maximal overlap wavelet transform (MODWT) of a time series, so we start with a discussion of this transform and its basic properties. The MODWT is closely related to transforms with a variety of names in the literature, including "undecimated DWT" (Shensa (1992) implemented via the "à trous algorithm"), "shift invariant DWT" (Beylkin, 1992; Lang et al., 1995), "wavelet frames" (Unser, 1995), "translation invariant DWT" (Coifman and Donoho, 1995; Liang and Parks, 1996; Del Marco and Weiss, 1997), "stationary DWT" (Nason and Silverman, 1995), "time invariant DWT" (Pesquet et al., 1996), and "nondecimated DWT" (Bruce and Gao, 1996). For more details about the MODWT, we refer the reader to Percival and Walden (2000), from which we adopt the notation used below.

The starting point for the MODWT is a Daubechies wavelet filter $\{\tilde{h}_{1,l}, l = 0, 1, \ldots, L_1 - 1\}$, where we insist that $\tilde{h}_{1,0} \neq 0$ and $\tilde{h}_{1,L_1-1} \neq 0$, so that the filter has width L_1 (for technical reasons, this width must be even). We define $\tilde{h}_{1,l} = 0$ for $l < 0$ and $l \geq L_1$ for convenience. By definition, a Daubechies wavelet filter must satisfy three properties as follows:

$$\sum_{l\in\mathbb{Z}} \tilde{h}_{1,l} = 0, \quad \sum_{l\in\mathbb{Z}} \tilde{h}_{1,l}^2 = 1/2, \quad \text{and} \quad \sum_{l\in\mathbb{Z}} \tilde{h}_{1,l}\tilde{h}_{1,l+2n} = 0, \quad n = \pm 1, \pm 2, \ldots, \tag{1}$$

where $\mathbb{Z}$ is the set of all integers. Although it is easy to construct filters that satisfy the first two properties, the third (orthogonality to even shifts) is challenging. The simplest filter with all three properties is the Haar wavelet filter, which has width $L_1 = 2$ and filter coefficients $\tilde{h}_{1,0} = 1/2$ and $\tilde{h}_{1,1} = -1/2$. We denote the transfer function (i.e., discrete Fourier transform (DFT)) for $\{\tilde{h}_{1,l}\}$ by

$$\widetilde{H}_1(f) \equiv \sum_{l\in\mathbb{Z}} \tilde{h}_{1,l} e^{-i2\pi f l}, \quad -\infty < f < \infty,$$

and its associated squared gain function by $\widetilde{\mathcal{H}}_1(f) \equiv |\widetilde{H}_1(f)|^2$. For the Haar wavelet filter, we have

$$\widetilde{H}_1(f) = \tfrac{1}{2} - \tfrac{1}{2} e^{-i2\pi f} \quad \text{and} \quad \widetilde{\mathcal{H}}_1(f) = \sin^2(\pi f). \tag{2}$$

The wavelet filter spawns a complementary filter known as the scaling filter, defined by

$$\tilde{g}_{1,l} = (-1)^{l+1} \tilde{h}_{1,L_1-l-1}, \quad l \in \mathbb{Z}.$$

In the Haar case, we have $\tilde{g}_{1,0} = 1/2$, $\tilde{g}_{1,1} = 1/2$, and $\tilde{g}_{1,l} = 0$ otherwise. We denote the corresponding transfer and squared gain functions by $\widetilde{G}_1(\cdot)$ and $\widetilde{\mathcal{G}}_1(\cdot)$. For the Haar scaling filter, we have

$$\widetilde{G}_1(f) = \tfrac{1}{2} + \tfrac{1}{2} e^{-i2\pi f} \quad \text{and} \quad \widetilde{\mathcal{G}}_1(f) = \cos^2(\pi f).$$

Although the wavelet filter is a high-pass filter with a nominal passband defined by $1/4 \le |f| \le 1/2$, the scaling filter is a low-pass filter with passband dictated by $0 \le |f| \le 1/4$. A fundamental (but far from obvious) consequence of conditions (1) is that the squared gain functions for the wavelet and scaling filters must satisfy

$$\widetilde{\mathcal{H}}_1(f) + \widetilde{\mathcal{G}}_1(f) = 1 \quad \text{for all } f. \tag{3}$$

An implication of the above equation is that applying both filters to a time series results in outputs that preserve the content of the original series over all Fourier frequencies. Note that the above relationship holds in the Haar case because of the well-known identity $\sin^2(x) + \cos^2(x) = 1$.

Let $\{X_t, t = 0, 1, \ldots, N-1\}$ represent a time series of N observations regularly sampled in time; that is, the time associated with X_t is $t_0 + t\Delta$, where t_0 is the time at which X_0 is observed, and Δ is the sampling interval between adjacent observations. Upon circularly filtering $\{X_t\}$ with the wavelet and scaling filters, we obtain the unit-level MODWT wavelet coefficients

$$\widetilde{W}_{1,t} \equiv \sum_{l=0}^{L_1-1} \tilde{h}_{1,l} X_{t-l \bmod N}, \quad t = 0, 1, \ldots, N-1,$$

and corresponding scaling coefficients

$$\widetilde{V}_{1,t} \equiv \sum_{l=0}^{L_1-1} \tilde{g}_{1,l} X_{t-l \bmod N}, \quad t = 0, 1, \ldots, N-1.$$

The modulo operator in the above equation is such that '$t - l \bmod N$' is equal to $t - l$ if $0 \le t - l \le N - 1$; otherwise, it is equal to $t - l + nN$, where n is the unique integer such that $0 \le t - l + nN \le N - 1$. This operator in effect ties the beginning and end of the time series together, which is why the filtering is referred to as circular. For the Haar case, we have

$$\widetilde{W}_{1,t} = \frac{X_t - X_{t-1}}{2} \quad \text{and} \quad \widetilde{V}_{1,t} = \frac{X_t + X_{t-1}}{2} \quad \text{for} \quad t = 1, 2, \ldots N-1,$$

whereas

$$\widetilde{W}_{1,0} = \frac{X_0 - X_{N-1}}{2} \quad \text{and} \quad \widetilde{V}_{1,0} = \frac{X_0 + X_{N-1}}{2}. \tag{4}$$

With the exception of $t = 0$, each scaling coefficient is the average of two adjacent values from the time series. We associate these averages with a standardized scale $\lambda_1 = 2$ and a physical scale of $\lambda_1 \Delta$. By contrast, each wavelet coefficient is proportional to the difference between two adjacent values. If we take the point of view that each X_t spans the time interval Δ (as would be appropriate if X_t were to represent, e.g., the average annual temperature at a particular spot on the earth so that $\Delta = 1$ year), then we can regard each wavelet coefficient as being proportional to the difference of averages over a standardized scale of $\tau_1 = 1$ and a physical scale of $\tau_1 \Delta$.

We can also interpret the MODWT scaling and wavelet coefficients as averages and differences between adjacently located averages when the transform is based on other Daubechies wavelet filters besides the Haar. For these other filters, each $\widetilde{V}_{1,t}$ at indices $t = L_1 - 1, \ldots, N-1$ is related to a *weighted* localized average of the time series over a scale of 2Δ, where we now take 2Δ to be a measure of the effective width of the weighted average. The wavelet coefficients $\widetilde{W}_{1,t}$ are related to changes in adjacent *weighted* localized averages over a scale of Δ. Our ability to make these interpretations requires that $\{\tilde{h}_l\}$ satisfies certain conditions above and beyond those imposed by Eq. (1). For example, the regularity conditions that lead to $\{\tilde{h}_{1,l}\}$ having a squared gain function of

$$\widetilde{\mathcal{H}}_1(f) = \sin^{L_1}(\pi f) \sum_{l=0}^{\frac{L_1}{2}-1} \binom{\frac{L_1}{2} - 1 + l}{l} \cos^{2l}(\pi f) \tag{5}$$

allow us to attach these interpretations to $\widetilde{V}_{1,t}$ and $\widetilde{W}_{1,t}$. Note that the above equation collapses to Eq. (2) in the Haar case ($L_1 = 2$). The squared gain functions for the widely used "least asymmetric" Daubechies wavelet filters take the above form.

It is important to note that the wavelet and scaling coefficients with indices $t = 0, 1, \ldots, L_1 - 2$ do *not* involve localized averages. Instead, they combine values from both the beginning and end of the time series. We refer to these special cases as "boundary" coefficients, to which we will need to pay special attention. In the Haar case, the unit-level boundary coefficients are shown in Eq. (4).

Just as the unit-level wavelet coefficients are related to differences of averages at scale $\tau_1 = 1$ while the scaling coefficients extract averages from $\{X_t\}$ at scale $\lambda_1 = 2$, higher-level MODWT coefficients extract quantities with similar interpretations for larger scales $\tau_j = 2^{j-1}$ and $\lambda_j = 2^j$, where j is the level index. We define the level $j > 1$ coefficients in terms of the higher-level wavelet filter $\{\tilde{h}_{j,l}, l = 0, 1, \ldots, L_j - 1\}$ and scaling filter $\{\tilde{g}_{j,l}, l = 0, 1, \ldots, L_j - 1\}$, where $L_j \equiv (2^j - 1)(L_1 - 1) + 1$. The appropriate definitions are

$$\widetilde{W}_{j,t} \equiv \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l \bmod N}, \quad t = 0, 1, \ldots, N-1 \tag{6}$$

and

$$\widetilde{V}_{j,t} \equiv \sum_{l=0}^{L_{J_0}-1} \tilde{g}_{j,l} X_{t-l \bmod N}, \quad t = 0, 1, \ldots, N-1. \tag{7}$$

For the Haar case, we have

$$\widetilde{W}_{j,t} = \frac{1}{2^j}\left(\sum_{l=0}^{2^{j-1}-1} X_{t-l} - \sum_{l=0}^{2^{j-1}-1} X_{t-l-2^{j-1}} \right) \quad \text{and} \quad \widetilde{V}_{j,t} = \frac{1}{2^j} \sum_{l=0}^{2^j-1} X_{t-l}$$

for $t = 2^{j-1}, 2^{j-1} + 1, \ldots N-1$. The scaling coefficients are averages over scale $\lambda_j = 2^j$, whereas the wavelet coefficients are proportional to differences of adjacent

averages over scale $\tau_j = 2^{j-1}$. For wavelets other than the Haar, the higher-level filters depend on just the basic wavelet and scaling filters $\{h_{1,l}\}$ and $\{g_{1,l}\}$ and are most easily described in terms of inverse DFTs of their transfer functions. The transfer functions for $\{\tilde{h}_{j,l}\}$ and $\{\tilde{g}_{j,l}\}$ are given by

$$\widetilde{H}_j(f) \equiv \widetilde{H}_1(2^{j-1}f) \prod_{l=0}^{j-2} \widetilde{G}_1(2^l f) \quad \text{and} \quad \widetilde{G}_j(f) \equiv \prod_{l=0}^{j-1} \widetilde{G}_1(2^l f).$$

The higher-level wavelet filters are band-pass filters with nominal passbands dictated by $1/2^{j+1} \le |f| \le 1/2^j$, whereas $\{\tilde{g}_{j,l}\}$ is a low-pass filter with passband given by $0 \le |f| \le 1/2^{j+1}$. For future use, we let

$$\widetilde{\mathcal{H}}_j(f) \equiv |\widetilde{H}_j(f)|^2 \quad \text{and} \quad \widetilde{\mathcal{G}}_j(f) \equiv |\widetilde{G}_j(f)|^2 \tag{8}$$

denote the corresponding squared gain functions.

Finally, we note that in practice, the MODWT wavelet and scaling coefficients are not computed directly via (6) and (7), but rather via an efficient recursive procedure known as the pyramid algorithm (for pseudo-code describing this algorithm, see Percival and Walden, 2000, pp. 177–178)).

3. Analysis of variance via the MODWT

Let $\boldsymbol{X}$, $\widetilde{\boldsymbol{W}}_j$, and $\widetilde{\boldsymbol{V}}_j$ be column vectors of dimension N whose tth elements are, respectively, X_t, $\widetilde{W}_{1,t}$, and $\widetilde{V}_{1,t}$. Let

$$\|\boldsymbol{X}\|^2 \equiv \sum_{t=0}^{N-1} X_t^2$$

be the square of the Euclidean norm of $\boldsymbol{X}$. We refer to $\|\boldsymbol{X}\|^2$ as the "energy" in $\boldsymbol{X}$. A key point about the MODWT is that it is energy preserving, in the sense that

$$\|\boldsymbol{X}\|^2 = \|\widetilde{\boldsymbol{W}}_1\|^2 + \|\widetilde{\boldsymbol{V}}_1\|^2. \tag{9}$$

In general, this decomposition of the energy into two parts follows from Parseval's theorem and Eq. (3). In the Haar case, it readily follows from

$$\widetilde{W}_{1,t}^2 + \widetilde{V}_{1,t}^2 = \frac{(X_t + X_{t-1})^2}{4} + \frac{(X_t - X_{t-1})^2}{4} = \frac{X_t^2 + X_{t-1}^2}{2},$$

$t = 1, \ldots, N-1$, along with a similar piece involving the boundary coefficients $\widetilde{W}_{1,0}$ and $\widetilde{V}_{1,0}$.

Letting $\overline{X} = \sum_{t=0}^{N-1} X_t/N$ represent the sample mean of $\boldsymbol{X}$, we can express the sample variance of our time series as

$$\hat{\sigma}_X^2 \equiv \frac{1}{N} \sum_{t=0}^{N-1} \left(X_t - \overline{X}\right)^2 = \frac{1}{N}\|\boldsymbol{X}\|^2 - \overline{X}^2 = \frac{1}{N}\|\widetilde{\boldsymbol{W}}_1\|^2 + \left(\frac{1}{N}\|\widetilde{\boldsymbol{V}}_1\|^2 - \overline{X}^2\right)$$
$$\equiv \hat{\sigma}_{W_1}^2 + \hat{\sigma}_{V_1}^2,$$

where $\hat{\sigma}^2_{W_1}$ and $\hat{\sigma}^2_{V_1}$ can be taken to be sample variances for $\widetilde{\boldsymbol{W}}_1$ and $\widetilde{\boldsymbol{V}}_1$ (the definition of the wavelet filter ensures that, if the X_ts have a population mean, then the population mean of the $\widetilde{W}_{1,t}$s is zero under mild conditions; on the other hand, the sample mean of $\widetilde{\boldsymbol{V}}_1$ is always $\overline{X}$, as is easy to verify directly in the Haar case). Thus, we can break up the sample variance of $\boldsymbol{X}$ into two parts, one of which ($\hat{\sigma}^2_{W_1}$) is attributable to changes in the time series over standardized scale $\tau_1 = 1$, and the other ($\hat{\sigma}^2_{V_1}$), to averages in $\boldsymbol{X}$ over scale $\lambda_1 = 2\tau_1 = 2$; alternatively, we can think of $\hat{\sigma}^2_{W_1}$ and $\hat{\sigma}^2_{V_1}$ as capturing the parts of $\hat{\sigma}^2_X$ due to high- and low-frequency fluctuations, respectively.

We can generalize the above scheme to define ANOVAs out to some maximum level $J_0 \geq 1$. Considering a level $J_0 = 2$ ANOVA, first, the basic idea is to replace $\|\widetilde{\boldsymbol{V}}_1\|^2$ in Eq. (9) with the sum of two values, namely, $\|\widetilde{\boldsymbol{W}}_2\|^2$ and $\|\widetilde{\boldsymbol{V}}_2\|^2$, the first of which is related to changes in adjacent weighted localized averages of $\{X_t\}$ over a scale of $\tau_2 = 2$, and the second, to weighted localized averages over a scale of $\lambda_2 = 2\tau_2 = 4$. By recursively replacing $\|\widetilde{\boldsymbol{V}}_{j-1}\|^2$ with $\|\widetilde{\boldsymbol{W}}_j\|^2 + \|\widetilde{\boldsymbol{V}}_j\|^2$, we are led to the level J_0 decomposition:

$$\|\boldsymbol{X}\|^2 = \sum_{j=1}^{J_0} \|\widetilde{\boldsymbol{W}}_j\|^2 + \|\widetilde{\boldsymbol{V}}_{J_0}\|^2 \quad \text{and} \quad \hat{\sigma}^2_X = \sum_{j=1}^{J_0} \hat{\sigma}^2_{W_j} + \hat{\sigma}^2_{V_{J_0}},$$

where

$$\hat{\sigma}^2_{W_j} \equiv \frac{1}{N}\|\widetilde{\boldsymbol{W}}_j\|^2 \quad \text{and} \quad \hat{\sigma}^2_{V_{J_0}} \equiv \frac{1}{N}\|\widetilde{\boldsymbol{V}}_{J_0}\|^2 - \overline{X}^2. \tag{10}$$

We refer to $\hat{\sigma}^2_{W_j}$ as the jth level empirical wavelet variance and to $\sigma^2_{V_{J_0}}$ as the level J_0 empirical scaling variance. The interpretation of $\hat{\sigma}^2_{W_j}$ is that it is related to changes in adjacent weighted localized averages of $\{X_t\}$ over a scale of $\tau_j = 2^{j-1}$, while $\sigma^2_{V_{J_0}}$ is associated with weighted localized averages over a scale of $\lambda_{J_0} = 2^{J_0}$. This generalization of Eq. (9) again follows from Parseval's theorem, but this time in conjunction with a generalization of Eq. (3), namely,

$$\sum_{j=1}^{J_0} \widetilde{\mathcal{H}}_j(f) + \widetilde{\mathcal{G}}_{J_0}(f) = 1 \ \text{ for all } f. \tag{11}$$

As a simple example of a wavelet-based ANOVA, consider a small segment of length $N = 192$ from a time series of subtidal sea-level fluctuations (Fig. 1a); see Percival and Mofjeld (1997) for details about these data. This segment is of interest because of several bumps, each spanning approximately 16 units of time. Fig. 1b shows the empirical wavelet variances $\hat{\sigma}^2_{W_j}$ (circles) and empirical scaling variance $\hat{\sigma}^2_{V_{J_0}}$ (asterisk) based on a level $J_0 = 7$ Haar MODWT. The sum of these eight variances is exactly equal to the sample variance $\hat{\sigma}^2_X \doteq 258.6$ of the time series. The largest wavelet variance occurs at level $j = 5$, which corresponds to scale $\tau_5 = 2^4 = 16$. The peak at this scale quantifies what a visual inspection picks out, namely, features in the series (the bumps) with a characteristic span of 16 time units. The fact that the scaling variance

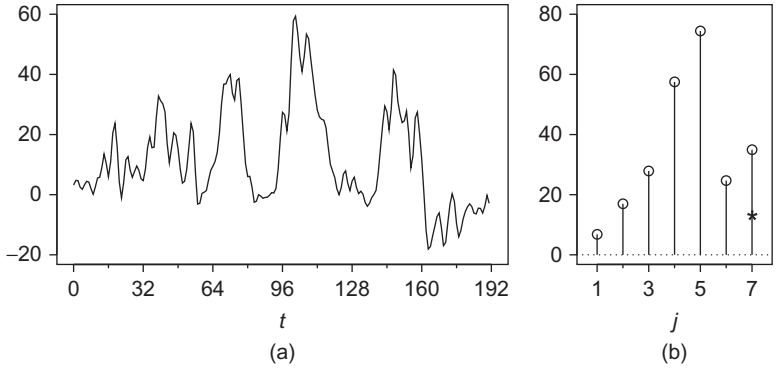

Fig. 1. Subtidal sea-level time series (a) and the Haar empirical wavelet variances (b) at levels $j = 1, \ldots, 7$ (circles) and empirical scaling variance for level 7 (asterisk). The seven levels correspond to scales $\tau_j = 2^{j-1}$, so the peak at level $j = 5$ is associated with changes on scale $\tau_5 = 16$. (Reflection boundary conditions were used in forming the MODWT – see Section 5.2 for details.)

at level $J_0 = 7$ is small relative to the displayed wavelet variances tells us that the bulk of the variance of the time series can be attributed to changes in averages over scales $\tau_7 = 2^6 = 64$ and smaller. Thus, the scale-based ANOVA given by the wavelet variance can offer an intuitively sensible explanation of how a time series is structured.

4. Definition and basic properties of wavelet variance

In this section, we formulate the wavelet variance for a dth order intrinsically stationary process $\{X_t : t \in \mathbb{Z}\}$, where $d \geq 0$ is an integer. The definition for such a process is as follows. If $d = 0$, then $\{X_t\}$ is just a second-order stationary process; that is, its expected value $E\{X_t\}$ and covariances $\operatorname{cov}\{X_{t+\tau}, X_t\}$, $\tau \in \mathbb{Z}$, are finite and do not depend on t. For $d > 0$, let $\{X_t^{(d)}\}$ represent the dth order backward difference of $\{X_t\}$:

$$X_t^{(d)} \equiv \sum_{k=0}^{d} \binom{d}{k} (-1)^k X_{t-k}$$

(thus $X_t^{(1)} = X_t - X_{t-1}$, $X_t^{(2)} = X_t - 2X_{t-1} + X_{t-2}$ and so forth). Then $\{X_t^{(d)}\}$ is second-order stationary, but $\{X_t^{(d-1)}\}$, $\{X_t^{(d-2)}\}$, …, $\{X_t^{(0)}\}$ are not, where we define $X_t^{(0)} = X_t$ for convenience. For example, suppose $\{Z_t\}$ is a white noise process, i.e., a sequence of uncorrelated random variables (RVs) with zero mean and finite variance. Then the nonstationary random walk process $X_t = \sum_{u=0}^{t} Z_t$ for $t \geq 1$ is first-order intrinsically stationary since its first difference is a stationary process. As a second example, suppose a and $b \neq 0$ are constants, and now let $X_t = a + bt + Z_t$ so that $\{X_t\}$ is nonstationary. Then $X_t^{(1)} = b + Z_t - Z_{t-1}$ is stationary, so $\{X_t\}$ is first-order intrinsically stationary. A special case of a dth-order intrinsically stationary process is an ARIMA(p, d, q) process, which is the most widely used parametric model in time series analysis (see, e.g., Brockwell and Davis (2002)).

For use below, let $\{s_\tau^{(d)} : \tau \in \mathbb{Z}\}$ denote the autocovariance sequence (ACVS) for $\{X_t^{(d)}\}$:

$$s_\tau^{(d)} \equiv \operatorname{cov}\left\{X_{t+\tau}^{(d)}, X_t^{(d)}\right\} = E\left\{(X_{t+\tau}^{(d)} - \mu^{(d)})(X_t^{(d)} - \mu^{(d)})\right\},$$

where $\mu^{(d)} \equiv E\{X_t^{(d)}\}$. Under the assumption that $\{X_t^{(d)}\}$ has an SDF $S_{X^{(d)}}(\cdot)$, we can, when $d > 0$, define a generalized SDF for $\{X_t\}$ itself via

$$S_X(f) = \frac{S_{X^{(d)}}(f)}{[4\sin^2(\pi f)]^d} \tag{12}$$

(Yaglom 1958); here, $4\sin^2(\pi f)$ comes into play because it defines the squared gain function for a first-order backward difference filter (cf. Eq. (2)).

Given a dth-order intrinsically stationary process $\{X_t : t \in \mathbb{Z}\}$ with an SDF given by Eq. (12) and a jth-level wavelet filter $\{\tilde{h}_{j,l}, l = 0, 1, \ldots, L_j - 1\}$, we can define an associated wavelet coefficient process via

$$\overline{W}_{j,t} \equiv \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l}. \tag{13}$$

Under the assumptions that the associated unit-level wavelet filter has a squared gain function given by Eq. (5) and that $L_1 \geq 2d$, the wavelet coefficient process is stationary with an SDF given by

$$S_j(f) \equiv \widetilde{\mathcal{H}}_j(f) S_X(f) = \frac{\widetilde{\mathcal{H}}_j(f) S_{X^{(d)}}(f)}{[4\sin^2(\pi f)]^d},$$

where $\widetilde{\mathcal{H}}_j(\cdot)$ is defined by Eq. (8). The squared gain function $\widetilde{\mathcal{H}}_j(\cdot)$ depends on $\widetilde{\mathcal{H}}_1(\cdot)$ of Eq. (5), which can be interpreted as arising from an implicit cascade of two filters. The first is a backward difference filter of order $L_1/2$. The condition $L_1 \geq 2d$ thus ensures that the jth-level wavelet filter has enough embedded differencing operations to transform $\{X_t\}$ into the stationary process $\{X_t^{(d)}\}$ (the second filter in the cascade transforms $\{X_t^{(d)}\}$ into $\overline{W}_{j,t}$). The jth-level wavelet variance is just the variance of the stationary process $\{\overline{W}_{j,t}\}$:

$$\nu_X^2(\tau_j) \equiv \operatorname{var}\{\overline{W}_{j,t}\} = \int_{-1/2}^{1/2} S_j(f) df.$$

If $d = 0$ so that $\{X_t\}$ is stationary, then

$$\sum_{j=1}^{\infty} \nu_X^2(\tau_j) = \operatorname{var}\{X_t\},$$

and hence the wavelet variance is a scale-based ANOVA for $\{X_t\}$, paralleling the empirical ANOVA for the sample variance described in Section 3. If $d > 0$, the

summation above diverges to infinity, which is a reasonable definition for the variance of certain (but not all) nonstationary processes with stationary differences.

For $d \geq 0$, the wavelet variance just depends on the ACVS $\{s_\tau^{(d)}\}$ for the stationary process $\{X_t^{(d)}\}$ at the heart of $\{X_t\}$:

$$\begin{aligned}\nu_X^2(\tau_j) &= \sum_{l=0}^{L_j-d-1} \sum_{m=0}^{L_j-d-1} \tilde{b}_{j,l}^{(d)} \tilde{b}_{j,m}^{(d)} s_{l-m}^{(d)} \\ &= s_0^{(d)} \sum_{l=0}^{L_j-d-1} \left(\tilde{b}_{j,l}^{(d)}\right)^2 + 2 \sum_{\tau=1}^{L_j-d-1} s_\tau^{(d)} \sum_{l=0}^{L_j-d-1-\tau} \tilde{b}_{j,l}^{(d)} \tilde{b}_{j,l+\tau}^{(d)}, \end{aligned} \tag{14}$$

where $\{\tilde{b}_{j,l}^{(d)}\}$ is the dth-order cumulative summation of $\{\tilde{h}_{j,l}\}$; that is, with $\tilde{b}_{j,l}^{(0)} \equiv \tilde{h}_{j,l}$, we have, for $k = 1, \ldots, d$,

$$\tilde{b}_{j,l}^{(k)} = \sum_{n=0}^{l} \tilde{b}_{j,n}^{(k-1)}, \quad l = 0, 1, \ldots, L_j - k - 1$$

(Lemma 1; Craigmile and Percival, 2005). This expression for $\nu_X^2(\tau_j)$ follows from reexpressing Eq. (13) in terms of $\{\tilde{b}_{j,l}^{(d)}\}$ and $\{X_t^{(d)}\}$:

$$\overline{W}_{j,t} = \sum_{l=0}^{L_j-d-1} \tilde{b}_{j,l}^{(d)} X_{t-l}^{(d)}. \tag{15}$$

The above equation directly leads to an expression for the ACVS for $\{\overline{W}_{j,t}\}$, namely,

$$s_{j,\tau} \equiv \operatorname{cov}\{\overline{W}_{j,t+\tau}, \overline{W}_{j,t}\} = \sum_{l=0}^{L_j-d-1} \sum_{m=0}^{L_j-d-1} \tilde{b}_{j,l}^{(d)} \tilde{b}_{j,m}^{(d)} s_{\tau+l-m}^{(d)}.$$

Eq. (14) follows from the above since $\nu_X^2(\tau_j) = s_{j,0}$.

When $d = 0$ or 1, we can also express the wavelet variance in terms of the semivariogram, defined as $\gamma_\tau = \frac{1}{2} \operatorname{var}\{X_\tau - X_0\}$. We then have

$$\nu_X^2(\tau_j) = -\sum_{l=0}^{L_j-1} \sum_{m=0}^{L_j-1} \tilde{h}_{j,l} \tilde{h}_{j,m} \gamma_{l-m}.$$

5. Basic estimators of the wavelet variance

We now consider the problem of estimating the wavelet variance given a time series that can be regarded as a realization of a portion $X_0, X_1, \ldots, X_{N-1}$ of a dth-order intrinsically stationary process. Under the assumption that $L_1 \geq 2d$, we can base an

estimator for $\nu_X^2(\tau_j)$ on the level j MODWT wavelet coefficients $\widetilde{W}_{j,t}$ of Eq. (6). If we compare

$$\widetilde{W}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l \bmod N} \quad \text{with} \quad \overline{W}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l} \quad \text{for} \quad t = 0, 1, \ldots, N-1,$$

we see that $\widetilde{W}_{j,t} = \overline{W}_{j,t}$ when $t \geq L_j - 1$, but this equality need not hold when $0 \leq t < L_j - 1$, that is, when $\widetilde{W}_{j,t}$ is a boundary coefficient. As discussed in the next two subsections, excluding the boundary coefficients leads us to an unbiased estimator of the wavelet variance, whereas, with certain modifications, we can form an attractive biased estimator that makes use of all available coefficients.

5.1. *Unbiased estimators of the wavelet variance*

Under the assumption that $E\{\overline{W}_{j,t}\} = 0$ and that $M_j \equiv N - L_j + 1 > 0$, an unbiased estimator of $\nu_X^2(\tau_j)$ is given by

$$\hat{\nu}_X^2(\tau_j) \equiv \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \widetilde{W}_{j,t}^2 = \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \overline{W}_{j,t}^2. \tag{16}$$

An unbiased estimator is not possible in general without the condition $E\{\overline{W}_{j,t}\} = 0$. As can be seen from Eq. (15), this condition will hold if $E\{X_t^{(d)}\} = 0$, which in general cannot be guaranteed; however, no matter what $E\{X_t^{(d)}\}$ is, the condition will hold as long as

$$\sum_{l=0}^{L_j-d-1} \tilde{b}_{j,l}^{(d)} = 0,$$

which we *can* guarantee by assuming $L_1 > 2d$ rather than just $L_1 \geq 2d$. The zero mean condition is thus easy to achieve by just increasing the length of the basic wavelet filter.

The large sample distribution for $\hat{\nu}_X^2(\tau_j)$ is tractable if we make some additional assumptions about the wavelet coefficient process. One pathway is to assume that $\{\overline{W}_{j,t}\}$ is a stationary Gaussian (normal) process with a square summable ACVS, that is,

$$A_j \equiv \sum_{\tau=-\infty}^{\infty} s_j^2 < \infty. \tag{17}$$

With this assumption, it follows that

$$\frac{M_j^{1/2}(\hat{\nu}_X^2(\tau_j) - \nu_X^2(\tau_j))}{(2A_j)^{1/2}} \stackrel{\mathrm{d}}{=} \mathcal{N}(0,1) \tag{18}$$

asymptotically, where "$\stackrel{\mathrm{d}}{=}$" stands for "is equal in distribution to", and $\mathcal{N}(0,1)$ is a standard Gaussian RV (Mondal (2007) gives a succinct proof of the above based on Theorem 5 of Giraitis and Surgailis (1985) and discusses earlier – but more complicated and less general – proofs in Percival (1983) for the Haar wavelet and in Percival (1995) for general Daubechies wavelet filters). Accordingly, the random interval

$$\left[\hat{\nu}_X^2(\tau_j) - \Phi^{-1}(1-p)\left(\frac{2A_j}{M_j}\right)^{1/2},\ \hat{\nu}_X^2(\tau_j) + \Phi^{-1}(1-p)\left(\frac{2A_j}{M_j}\right)^{1/2}\right] \tag{19}$$

constitutes an approximate $100(1-2p)\%$ confidence interval (CI) for $\nu_X^2(\tau_j)$, where $\Phi^{-1}(p)$ is the $p \times 100\%$ point for the standard Gaussian distribution.

The lower limit of the CI displayed in (19) is not restricted to be positive even though the true wavelet variance is. This fact poses a problem if we adopt the common practice of plotting estimates $\hat{\nu}_X^2(\tau_j)$ and their associated CIs on a logarithmic scale. An alternative – but asymptotically equivalent – approach that yields CIs with positive lower limits is to assume that asymptotically

$$\frac{\eta_j \hat{\nu}_X^2(\tau_j)}{\nu_X^2(\tau_j)} \stackrel{\mathrm{d}}{=} \chi_{\eta_j}^2, \tag{20}$$

where $\chi_{\eta_j}^2$ is a chi-square RV with η_j degrees of freedom. We can set η_j using a moment-matching scheme. Recalling first that $E\{\chi_{\eta_j}^2\} = \eta_j$ and $\mathrm{var}\{\chi_{\eta_j}^2\} = 2\eta_j$ so that

$$\frac{2\left(E\left\{c\chi_{\eta_j}^2\right\}\right)^2}{\mathrm{var}\left\{c\chi_{\eta_j}^2\right\}} = \eta_j \quad \text{for any constant } c,$$

we use the facts that $E\{\hat{\nu}_X^2(\tau_j)\} = \nu_X^2(\tau_j)$ and $\mathrm{var}\{\hat{\nu}_X^2(\tau_j)\} \approx 2A_j/M_j$ to obtain

$$\eta_j = \frac{2\left(E\{\hat{\nu}_X^2(\tau_j)\}\right)^2}{\mathrm{var}\{\hat{\nu}_X^2(\tau_j)\}} = \frac{2\nu_X^4(\tau_j)}{\mathrm{var}\{\hat{\nu}_X^2(\tau_j)\}} \approx \frac{M_j \nu_X^4(\tau_j)}{A_j}. \tag{21}$$

The random interval

$$\left[\frac{\eta_j \hat{\nu}_X^2(\tau_j)}{Q_{\eta_j}(1-p)}, \frac{\eta_j \hat{\nu}_X^2(\tau_j)}{Q_{\eta_j}(p)}\right] \tag{22}$$

is then an approximate $100(1-2p)\%$ CI for $\nu_X^2(\tau_j)$, where $Q_{\eta_j}(p)$ is the $p \times 100\%$ point for the $\chi_{\eta_j}^2$ distribution.

We must know A_j to form the CIs of Eqs (19) and (22). If we regard $\widetilde{W}_{j,L_j-1}, \ldots, \widetilde{W}_{j,N-1}$ as a time series whose mean value is zero, then we can estimate its ACVS via

$$\hat{s}_{j,\tau} \equiv \frac{1}{M_j} \sum_{t=L_j-1}^{N-|\tau|-1} \widetilde{W}_{j,t} \widetilde{W}_{j,t+|\tau|}, \quad 0 \le |\tau| \le M_j - 1.$$

An approximately unbiased estimator of A_j is given by

$$\hat{A}_j \equiv \sum_{\tau=-(M_j-1)}^{M_j-1} \frac{\hat{s}^2_{j,\tau}}{2} = \frac{\hat{s}^2_{j,0}}{2} + \sum_{\tau=1}^{M_j-1} \hat{s}^2_{j,\tau} = \frac{\hat{\nu}^4(\tau_j)}{2} + \sum_{\tau=1}^{M_j-1} \hat{s}^2_{j,\tau} \tag{23}$$

(comparison of $\hat{A}_j$ with the definition of A_j in Eq. (17) indicates a counter-intuitive division by two in the above – this is due to the moments of the χ^2_2 distribution, as explained in the study by Percival and Walden (2000), Section 8.4). We can plug this estimator into Eq. (19) to get Gaussian-based approximate CIs. We can also plug $\hat{A}_j$ along with $\hat{\nu}^4(\tau_j)$ into Eq. (21) to estimate η_j, which in turn can be used in Eq. (22) to produce $\chi^2_{\eta_j}$-based approximate CIs. Monte Carlo studies indicate that CIs based on estimating A_j are reasonably accurate as long as $M_j \geq 128$. If there aren't enough wavelet coefficients to get a decent estimate of A_j, a fallback for getting an approximate CI for $\nu^2(\tau_j)$ is to use Eq. (22) with η_j set to $\max\{M_j/2^j, 1\}$. This approach banks on the fact that the wavelet filter $\{\tilde{h}_{j,l}\}$ is an approximate band-pass filter and hence that $\{\overline{W}_{j,t}\}$ should resemble a band-limited process. If the SDF for $\{\overline{W}_{j,t}\}$ is relatively flat over the passband, this alternative approach is viable, but tends to produce conservative CIs.

The large-sample theory described above is based on the assumption that $\{\widetilde{W}_{j,t}\}$ is Gaussian. If $\{X_t\}$ itself is Gaussian, then $\{\widetilde{W}_{j,t}\}$ must be Gaussian since each $\widetilde{W}_{j,t}$ is a linear combination of RVs in $\{X_t\}$. For certain non-Gaussian processes $\{X_t\}$, the assumption that $\{\widetilde{W}_{j,t}\}$ is approximately Gaussian is viable (particularly for large j) because linear filtering tends to induce Gaussianity (Mallows, 1967). If $\{\widetilde{W}_{j,t}\}$ cannot be regarded as approximately Gaussian, we can obtain a large sample approximation to the distribution of $\hat{\nu}^2(\tau_j)$ if we are willing to make certain assumptions (Serroukh et al., 2000). Let $\mathcal{M}^0_{-\infty}$ and $\mathcal{M}^\infty_n$ denote the σ-algebras generated by $\{\ldots, \widetilde{W}_{j,-1}, \widetilde{W}_{j,0}\}$ and $\{\widetilde{W}_{j,n}, \widetilde{W}_{j,n+1}, \ldots\}$, respectively. For $n > 0$, define the mixing coefficient

$$\alpha_n = \sup_{A \in \mathcal{M}^0_{-\infty}, B \in \mathcal{M}^\infty_n} |P(A \cap B) - P(A)P(B)|,$$

where $P(A)$ is the probability of the event A. If $\{\widetilde{W}_{j,t}\}$ is strictly stationary with $E\{|\widetilde{W}_{j,t}|^{4+2\delta}\} < \infty$ for some $\delta > 0$, if $\sum_{n=1}^\infty \alpha_n^{\delta/(2+\delta)} < \infty$, and if the SDF $S_{\widetilde{W}^2_{j,t}}(f)$ for the process $\{\widetilde{W}^2_{j,t}\}$ is positive at zero frequency, then

$$\frac{M_j^{1/2}(\hat{\nu}^2_X(\tau_j) - \nu^2_X(\tau_j))}{S^{1/2}_{\widetilde{W}^2_{j,t}}(0)} \stackrel{\mathrm{d}}{=} \mathcal{N}(0,1) \tag{24}$$

approximately for large M_j. The condition on the mixing coefficients implies that $\alpha_n \to 0$ as $n \to \infty$, in which case $\{\widetilde{W}_{j,t}\}$ is said to have strong mixing; that is, its dependence is short range (Rosenblatt, 1985, pp. 62–63). The assumptions we need for (24) to hold are thus restrictive but not overly so (Serroukh et al. (2000) give specific examples of processes for which these assumptions hold).

As in the Gaussian case, we can reexpress Eq. (24) in terms of a chi-square distribution as per Eq. (20), with the degrees of freedom now given by

$$\eta_j \approx \frac{2M_j \nu_X^4(\tau_j)}{S_{\widetilde{W}^2_{j,t}}(0)}. \tag{25}$$

To form the CI of Eq. (22), we need in practice to estimate $S_{\widetilde{W}^2_{j,t}}(0)$ based on $\widetilde{W}^2_{j,L_j-1}, \ldots, \widetilde{W}^2_{j,N-1}$. Serroukh et al. (2000) advocate a multitaper SDF estimator of order $K = 5$ based on Slepian (discrete prolate spheroidal) data tapers $\{v_{k,t}\}$, $k = 0, 1, \ldots, K-1$, with design bandwidth set to $7/M_j$ (Percival and Walden 1993; Thomson 1982). Define

$$J_k(0) = \sum_{t=L_j-1}^{N-1} v_{k,t} \widetilde{W}^2_{j,t}, \quad V_k(0) = \sum_{t=L_j-1}^{N-1} v_{k,t} \quad \text{and} \quad \check{\nu}^2_X(\tau_j) = \frac{\sum_{k=0}^{K-1} J_k(0) V_k(0)}{\sum_{k=0}^{K-1} V_k^2(0)}. \tag{26}$$

Then the required estimator takes the form

$$\hat{S}_{\widetilde{W}^2_{j,t}}(0) = \frac{1}{K} \sum_{k=0}^{K-1} \left(J_k(0) - V_k(0) \check{\nu}^2_X(\tau_j) \right)^2. \tag{27}$$

Two comments are in order. First, computation of the above can be simplified by noting that $V_k(0) = 0$ for odd k. Second, as suggested by its notation, we can regard $\check{\nu}^2_X(\tau_j)$ as an estimator of the wavelet variance alternative to $\hat{\nu}^2_X(\tau_j)$ (both estimators are unbiased, but $\check{\nu}^2_X(\tau_j)$ is not constrained to be non-negative, even though the exponent "2" would suggest otherwise).

5.2. *Biased estimators of the wavelet variance*

The unbiased estimator $\hat{\nu}^2_X(\tau_j)$ of the wavelet variance makes use of just the nonboundary wavelet coefficients from the MODWT. The number M_j of such coefficients decreases drastically as the level index j increases. For example, for a wavelet filter of width $L_1 = 8$ and a time series of length $N = 1024$, we find $M_j = 1017$, 1003, 975, 919, 807, 583, and 135 for levels $j = 1, \ldots, 7$, and there are no nonboundary coefficients at levels $j \geq 8$ (these correspond to scales $\tau_j \geq 128$). This fact motivates us to consider estimators of $\nu^2_X(\tau_j)$ that make use of the $N - \max\{M_j, 0\}$ boundary wavelet coefficients. One obvious candidate is

$$\hat{\sigma}^2_{W_j} \equiv \frac{1}{N} \sum_{t=0}^{N-1} \widetilde{W}^2_{j,t},$$

which we introduced in the context of forming a scaled-based ANOVA for a time series (see Eq. (10)). This estimator is in general biased because $E\{\widetilde{W}^2_{j,t}\}$ need not equal $\nu^2_X(\tau_j)$ when $t = 0, \ldots, L_j - 2$. If the underlying time series $\{X_t\}$ is a stationary process, then $E\{\hat{\sigma}^2_{W_j}\} \to \nu^2_X(\tau_j)$ as $N \to \infty$; however, if $\{X_t\}$ is first-order intrinsically

stationary, then $\hat{\sigma}^2_{W_j}$ is not in generally asymptotically unbiased. The basic reason is that $E\{(X_0 - X_{N-1})^2\} \to \infty$, which means that we can expect to see an increasing mismatch between the beginning and the end of the time series. This mismatch adversely impacts the boundary wavelet coefficients because they join together X_ts from both ends of the time series.

We can create a biased estimator that is asymptotically equivalent to $\hat{\nu}^2_X(\tau_j)$ for zeroth- and first-order intrinsically stationary processes by using wavelet coefficients obtained from an augmented version of our original time series $X_0, \ldots, X_{N-1}$. To do so, we append a time-reversed version of the series to the original X_ts to create a series of length $2N$:

$$X_0, X_1, \ldots, X_{N-2}, X_{N-1}, X_{N-1}, X_{N-2}, \ldots, X_1, X_0.$$

Denoting this series by $X'_0, \ldots, X'_{2N-1}$, we first note that its sample variance is identical to that for the original series $X_0, \ldots, X_{N-1}$, so that an ANOVA of the X'_ts is a useful surrogate for an ANOVA of the original X_ts. The MODWT of the X'_ts is given by

$$\widetilde{W}'_{j,t} \equiv \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X'_{t-l \bmod 2N}, \quad t = 0, 1, \ldots, 2N-1,$$

which we also refer to as the MODWT of the X_ts based on reflection boundary conditions. Greenhall et al. (1999) proposed the wavelet variance estimator

$$\overleftrightarrow{\nu}^2_X(\tau_j) \equiv \frac{1}{2N} \sum_{t=0}^{2N-1} (\widetilde{W}'_{j,t})^2,$$

but only explored its statistical properties through limited computer experiments. Aldrich (2005) showed that, although this estimator is generally biased, it is asymptotically equivalent to $\hat{\nu}^2_X(\tau_j)$ for zeroth- and first-order intrinsically stationary processes (but not for second order and higher). He also compared the mean squared errors of $\overleftrightarrow{\nu}^2_X(\tau_j)$ and $\hat{\nu}^2_X(\tau_j)$ for representative processes through both exact expressions and computer experiments and found the biased estimator to be superior to the unbiased estimator, particularly in cases where M_j is small relative to N.

6. Specialized estimators of the wavelet variance

The basic estimators of the wavelet variance discussed in the previous section are predicated on certain assumptions that might not hold in practical situations. Here we consider estimation of the wavelet variance when faced with departures from one of two assumptions. The first assumption is that our time series consists of N contiguous values; that is, there are no missing values in the time series. The second is that our series does not suffer from contamination unrelated to the process of interest (i.e., from corrupted observations).

6.1. *Estimation of wavelet variance for gappy time series*

Suppose we use an automatic measuring system to record the temperature every day at noon at some outdoor location, with the intent of eventually collecting a time series of N temperature measurements regularly sampled in time. In practice, we might not achieve this goal for a variety of reasons (power outages, sporadic instrumentation failure, vandalism, etc.) so that we end up with gaps (missing values) in the collected series. The wavelet variance estimators we discussed in Section 5 presume a regularly sampled series. If we want to use one of these estimators, we are faced with the task of filling in missing values. There are a variety of methods for doing so, ranging from the simple (using the sample mean of the existing observations as a surrogate for the missing ones) to the sophisticated (developing a statistical model for the time series and using the model to fill in the gaps via conditional expectations or a stochastic interpolation scheme). Interpolation can work well for certain gappy time series (particularly if both the number of gaps and the gap sizes are small) but can be problematic for others. Here we discuss two specialized wavelet variance estimators proposed by Mondal and Percival (2010) to handle gappy time series without having to resort to interpolation.

Let $\{\delta_t\}$ be a strictly stationary binary-valued process such that δ_t is 0 or 1 according to whether X_t is missing or present (we assume that $E\{\delta_t\} > 0$ and that $\{\delta_t\}$ and $\{X_t\}$ are independent). Define

$$\beta_k^{-1} = P(\delta_t = 1 \text{ and } \delta_{t+k} = 1),$$

and, for $0 \le l \le L_j - 1$ and $0 \le l' \le L_j - 1$, let

$$\hat{\beta}_{l,l'}^{-1} = \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \delta_{t-l}\delta_{t-l'},$$

which is an estimator of $\beta_{l-l'}^{-1}$. While $\beta_k^{-1} > 0$ necessarily, we must assume that $\hat{\beta}_{l,l'}^{-1} > 0$ for all l, l'. This assumption is restrictive because it might not hold for a time series with too many gaps (it does hold asymptotically almost surely). Define a covariance-type estimator by

$$\hat{u}_X^2(\tau_j) = \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \sum_{l=0}^{L_j-1} \sum_{l'=0}^{L_j-1} \tilde{h}_{j,l}\tilde{h}_{j,l'}\hat{\beta}_{l,l'} X_{t-l}X_{t-l'}\delta_{t-l}\delta_{t-l'} \tag{28}$$

and a semi-variogram-type estimator by

$$\hat{v}_X^2(\tau_j) = -\frac{1}{2M_j} \sum_{t=L_j-1}^{N-1} \sum_{l=0}^{L_j-1} \sum_{l'=0}^{L_j-1} \tilde{h}_{j,l}\tilde{h}_{j,l'}\hat{\beta}_{l,l'}(X_{t-l} - X_{t-l'})^2\delta_{t-l}\delta_{t-l'}. \tag{29}$$

Both $\hat{u}_X^2(\tau_j)$ and $\hat{v}_X^2(\tau_j)$ collapse to the usual unbiased estimator $\hat{\nu}_X^2(\tau_j)$ when $\delta_t = 1$ for all t (the gap-free case). In the presence of gaps, the expected value of both estimators is $\nu_X^2(\tau_j)$, but it should be carefully noted that neither estimator is guaranteed to be non-negative for a gappy realization of $X_0, \ldots, X_{N-1}$.

Let us now consider the large sample theory for $\hat{u}_X^2(\tau_j)$ and $\hat{v}_X^2(\tau_j)$. Mondal and Percival (2010) show that, if $\{X_t\}$ is a stationary Gaussian process whose SDF $S_X(\cdot)$ is square integrable and if $\{\delta_t\}$ satisfies certain technical conditions in addition to strict stationarity, then $\hat{u}_X^2(\tau_j)$ is asymptotically Gaussian with mean $\nu_X^2(\tau_j)$ and large sample variance given by $S_{U_{j,t}^2}(0)/M_j$, where the numerator is the SDF at zero frequency for the stationary process

$$U_{j,t}^2 \equiv \sum_{l=0}^{L_j-1} \sum_{l'=0}^{L_j-1} \tilde{h}_{j,l} \tilde{h}_{j,l'} \beta_{l-l'} X_{t-l} X_{t-l'} \delta_{t-l} \delta_{t-l'}.$$

This process has a mean of $\nu_X^2(\tau_j)$ and collapses to $\overline{W}_{j,t}^2$ in the gap-free case (again, note carefully that $U_{j,t}^2$ can be negative for certain realizations). If we let

$$\widetilde{U}_{j,t}^2 \equiv \sum_{l=0}^{L_j-1} \sum_{l'=0}^{L_j-1} \tilde{h}_{j,l} \tilde{h}_{j,l'} \hat{\beta}_{l,l'} X_{t-l} X_{t-l'} \delta_{t-l} \delta_{t-l'}, \quad t = L_j - 1, \ldots, N-1,$$

we can estimate $S_{U_{j,t}^2}(0)$ using the multitaper approach of Eqs (26) and (27) with $J_k(0)$ redefined to be $\sum_t v_{k,t} \widetilde{U}_{j,t}^2$. On the other hand, if $\{X_t\}$ is a zeroth- or first-order intrinsically stationary Gaussian process such that $\sin^2(\pi f) S_X(f)$ is square integrable and if we make the same assumptions as before about $\{\delta_t\}$, then $\hat{v}_X^2(\tau_j)$ is asymptotically Gaussian with mean $\nu_X^2(\tau_j)$ and large sample variance given by $S_{V_{j,t}^2}(0)/M_j$, which, as before, involves an SDF at zero frequency, but this time for the stationary process

$$V_{j,t}^2 \equiv -\frac{1}{2} \sum_{l=0}^{L_j-1} \sum_{l'=0}^{L_j-1} \tilde{h}_{j,l} \tilde{h}_{j,l'} \beta_{l-l'} (X_{t-l} - X_{t-l'})^2 \delta_{t-l} \delta_{t-l'}.$$

Again, the above process has a mean of $\nu_X^2(\tau_j)$, collapses to $\overline{W}_{j,t}^2$ in the gap-free case and can be negative for certain realizations. Letting

$$\widetilde{V}_{j,t}^2 \equiv \sum_{l=0}^{L_j-1} \sum_{l'=0}^{L_j-1} \tilde{h}_{j,l} \tilde{h}_{j,l'} \hat{\beta}_{l,l'} (X_{t-l} - X_{t-l'})^2 \delta_{t-l} \delta_{t-l'}, \quad t = L_j - 1, \ldots, N-1,$$

we can estimate $S_{V_{j,t}^2}(0)$ via Eqs (26) and (27) with $J_k(0)$ now redefined to be $\sum_t v_{k,t} \widetilde{V}_{j,t}^2$. Mondal and Percival (2010) note that the Gaussianity assumption on $\{X_t\}$ can be dropped, and both estimators will still have the same limiting distribution if we assume mixing conditions similar to what was needed to obtain the result stated in Eq. (24).

Both $\hat{u}_X^2(\tau_j)$ and $\hat{v}_X^2(\tau_j)$ can handle stationary processes, but $\hat{v}_X^2(\tau_j)$ also works for first-order intrinsically stationary processes. It might seem we could dispense with $\hat{u}_X^2(\tau_j)$ in favor of $\hat{v}_X^2(\tau_j)$; however, for certain – but not all – stationary processes, $\hat{u}_X^2(\tau_j)$ proves to be more efficient asymptotically than $\hat{v}_X^2(\tau_j)$ as measured by the ratio $S_{V_{j,t}^2}(0)/S_{U_{j,t}^2}(0)$. There is thus a role for both estimators. However, we note one important practical distinction between them. The semi-variogram-type estimator $\hat{v}_X^2(\tau_j)$ is

invariant if we add a constant to the observed time series, whereas the covariance-type estimator $\hat{u}_X^2(\tau_j)$ is not. Thus it is important to center a time series by subtracting off its sample mean prior to computing $\hat{u}_X^2(\tau_j)$.

6.2. *Robust estimation of wavelet variance*

The usual unbiased estimator of the wavelet variance is the sample mean of squared wavelet coefficients. In general, sample means as an estimator of a population mean are particularly sensitive to contamination, that is, a small number of large values that do not reflect the statistical properties of the underlying process of interest. This fact has motivated the quest for robust alternatives to sample means that perform better in the presence of contamination. A simple robust estimator of $\nu_X^2(\tau_j)$ is the sample median of $\widetilde{W}_{j,L_j-1}^2, \ldots, \widetilde{W}_{j,N-1}^2$ after an adjustment to take into account the difference between the population mean $\nu_X^2(\tau_j)$ and the population median of the $\widetilde{W}_{j,t}^2$s (Stoev et al., 2006). We can develop an appropriate statistical theory for a median-type estimator of the wavelet variance by considering

$$\widetilde{Q}_{j,t} \equiv \log(\widetilde{W}_{j,t}^2).$$

Because the log of the median of the $\widetilde{W}_{j,t}^2$s is the same as the median of the $\widetilde{Q}_{j,t}$s, a large sample theory based on the latter is pertinent for an estimator based on the sample median of the $\widetilde{W}_{j,t}^2$s. The advantage of the log transform is that it recasts the median-type estimator as a special case of the M-estimators pioneered by Huber (1964). These estimators work with location parameters, whereas $\nu_X^2(\tau_j)$ is a scale parameter, but one that can be recast as a location parameter via the log transform. Focusing on the case where $\{\widetilde{W}_{j,t}\}$ is Gaussian, it follows from Bartlett and Kendall (1946) that

$$E\{\widetilde{Q}_{j,t}\} = \log(\nu_X^2(\tau_j)) + \psi\left(\tfrac{1}{2}\right) + \log(2) \equiv \mu_j \quad \text{and} \quad \operatorname{var}\{\widetilde{Q}_{j,t}\} = \psi'\left(\tfrac{1}{2}\right) = \frac{\pi^2}{2},$$

where ψ and ψ' are the di- and tri-gamma functions. These facts allow us to write

$$\widetilde{Q}_{j,t} = \mu_j + \epsilon_{j,t}, \tag{30}$$

where $E\{\epsilon_{j,t}\} = 0$ and $\operatorname{var}\{\widetilde{Q}_{j,t}\} = \pi^2/2$. Thus we can manipulate a location estimator for $\widetilde{Q}_{j,t}$ so that it becomes an estimator of $\nu_X^2(\tau_j)$ since

$$\nu_X^2(\tau_j) = \exp(\mu_j - \psi\left(\tfrac{1}{2}\right) - \log(2)).$$

(We would need different manipulations to handle non-Gaussian processes.)

In general, an M-estimator for μ_j of Eq. (30) is based on a real-valued function $\varphi(\cdot)$ that is defined over the real axis $\mathbb{R}$ and satisfies certain technical conditions (see Mondal and Percival (2012a) for details). The M-estimator is

$$\hat{\mu}_j \equiv \arg\min_{x \in \mathbb{R}} \left| \sum_{t=L_j-1}^{N-1} \varphi(\widetilde{Q}_{j,t} - x) \right|.$$

Let us specialize to the case $\varphi(x) = \text{sign}(x)$, for which the M-estimator becomes the sample median of the $\widetilde{Q}_{j,t}$s. Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density and distribution functions for a standard Gaussian RV, and let $\Phi^{-1}(\cdot)$ be the inverse of $\Phi(\cdot)$. Under the assumption that $\{\widetilde{W}_{j,t}\}$ is a zero-mean Gaussian process with an SDF that is square integrable, Mondal and Percival (2012a) show that $\hat{\mu}_j$ is asymptotically Gaussian with mean

$$\mu_{0,j} = \log\left(\nu_X^2(\tau_j)\right) + 2\log\left(\Phi^{-1}\left(\tfrac{3}{4}\right)\right)$$

and large sample variance $S_\varphi(0)/(M_j C)$, which involves a constant $C = 4\left[\phi(\Phi^{-1}(\frac{3}{4}))\Phi^{-1}(\frac{3}{4})\right]^2$ and the SDF at zero frequency for the stationary process $\{\varphi(\widetilde{Q}_{j,t} - \mu_{0,j})\}$. We can estimate $S_\varphi(0)$ via the multitaper approach of Eqs (26) and (27) by redefining $J_k(0)$ to be $\sum_t v_{k,t}\varphi(\widetilde{Q}_{j,t} - \hat{\mu}_j)$. Denoting this estimator by $\hat{S}_\varphi(0)$, it can be shown that an approximately unbiased and robust estimator of $\nu_X^2(\tau_j)$ is given by

$$\hat{r}_X^2(\tau_j) = \frac{\text{median}\{\widetilde{W}_{j,t}^2\} \cdot \exp(-\hat{S}_\varphi(0)/[2M_j C])}{\left(\Phi^{-1}(\frac{3}{4})\right)^2}. \tag{31}$$

The above estimator is asymptotically normal with mean $\nu_X^2(\tau_j)$ and large sample variance $\nu_X^4(\tau_j)S_\varphi(0)/(M_j C)$. We can use this large sample theory to form CIs for $\nu_X^2(\tau_j)$ based on the median-type estimator $\hat{r}_X^2(\tau_j)$. (Mondal and Percival (2012a) provide theory paralleling the above for M-estimators other than the median.)

The median-type estimator $\hat{r}_X^2(\tau_j)$ guards against data contamination but is a less efficient estimator of $\nu_X^2(\tau_j)$ than the unbiased mean-type estimator $\hat{\nu}_X^2(\tau_j)$ when in fact the $\widetilde{W}_{j,t}$s are free of contamination. Mondal and Percival (2012a) found that, for moderate sample sizes, $\hat{r}_X^2(\tau_j)$ has approximately twice the variance of $\hat{\nu}_X^2(\tau_j)$ over a selection of stationary processes encompassing both short- and long-range dependence. Thus, if the $\widetilde{W}_{j,t}$s are truly Gaussian, we can expect $\hat{r}_X^2(\tau_j)$ to perform markedly poorer than $\hat{\nu}_X^2(\tau_j)$, but the presence of contamination can lead to the median-type estimator being preferred.

7. Combining wavelet variance estimators across scales

In the previous two sections, we have presented a variety of wavelet variance estimators, which, in conjunction with their sampling theory, can be used to form, say, 95% CIs for the true wavelet variance $\nu_X^2(\tau_j)$. Taking the unbiased estimator as an example, it is conventional to plot the estimates $\hat{\nu}_X^2(\tau_j)$ and associated CIs versus standardized scale τ_j (or physical scale $\tau_j\Delta$) on log/log axes, in part because the estimates and CIs can range over many orders of magnitude and in part because scales increase by factors of two as the level index j increases. In addition to telling us how the variance of a time series is partitioned out across different scales, two patterns warranting further analysis often emerge in plots of $\log(\hat{\nu}_X^2(\tau_j))$ versus $\log(\tau_j)$. The first pattern is a stretch of scales over which $\log(\hat{\nu}_X^2(\tau_j))$ varies approximately linearly with $\log(\tau_j)$. Two explanations for this linear pattern are that the intrinsically stationary process $\{X_t\}$ exhibits long-range dependence or fractal fluctuations, both of which are

special cases of a power-law variation. The second pattern is a peak at, say, scale τ_j; that is, we have both $\log\left(\hat{\nu}_X^2(\tau_j)\right) > \log\left(\hat{\nu}_X^2(\tau_{j-1})\right)$ and $\log\left(\hat{\nu}_X^2(\tau_j)\right) > \log\left(\hat{\nu}_X^2(\tau_{j+1})\right)$. Since the log transform preserves order, such a peak indicates a tendency of $\{X_t\}$ to have fluctuations over a so-called characteristic scale in the neighborhood of τ_j. In the subsections below, we look at methods for quantifying power-law variations and characteristic scales based on the wavelet variance. Both methods involve statistics that combine $\log\left(\hat{\nu}_X^2(\tau_j)\right)$ across adjacent scales. In preparation for delving into the sampling properties of these statistics, here we give some background on the statistical properties of $\log\left(\hat{\nu}_X^2(\tau_j)\right)$ (for simplicity, we focus on the unbiased estimator $\hat{\nu}_X^2(\tau_j)$, but the other estimators that we have discussed can be used instead with appropriate adjustments).

Under the assumption that $\hat{\nu}_X^2(\tau_j)$ obeys a chi-square distribution when properly normalized as per Eq. (20), we can write

$$\log\left(\hat{\nu}_X^2(\tau_j)\right) \stackrel{\mathrm{d}}{=} \log\left(\chi^2_{\eta_j}\right) + \log\left(\nu_X^2(\tau_j)\right) - \log(\eta_j).$$

Bartlett and Kendall (1946) show that

$$E\left\{\log(\chi^2_{\eta_j})\right\} = \psi\left(\tfrac{\eta_j}{2}\right) + \log(2),$$

where ψ is the di-gamma function. Hence, we have

$$E\left\{\log\left(\hat{\nu}_X^2(\tau_j)\right)\right\} = \log\left(\nu_X^2(\tau_j)\right) + \psi\left(\tfrac{\eta_j}{2}\right) + \log(2) - \log(\eta_j).$$

Assuming that the bivariate stationary processes $\{\widetilde{W}_{j,t}\}$ and $\{\widetilde{W}_{k,t}\}$ are jointly Gaussian with cross-covariance sequence $s_{j,k,\tau} \equiv \operatorname{cov}\{\widetilde{W}_{j,t+\tau}, \widetilde{W}_{k,t}\}$, we can approximate $\operatorname{cov}\{\log(\hat{\nu}_X^2(\tau_j)), \log(\hat{\nu}_X^2(\tau_k))\}$ by

$$\frac{\operatorname{cov}\left\{\hat{\nu}_X^2(\tau_j), \hat{\nu}_X^2(\tau_k)\right\}}{\nu_X^2(\tau_j)\nu_X^2(\tau_k)} + 2\frac{\operatorname{var}\left\{\hat{\nu}_X^2(\tau_j)\right\}\operatorname{var}\left\{\hat{\nu}_X^2(\tau_k)\right\} + \left(\operatorname{cov}\left\{\hat{\nu}_X^2(\tau_j), \hat{\nu}_X^2(\tau_k)\right\}\right)^2}{\nu_X^4(\tau_j)\nu_X^4(\tau_k)}, \quad (32)$$

where, for $j \le k$,

$$\operatorname{cov}\left\{\hat{\nu}_X^2(\tau_j), \hat{\nu}_X^2(\tau_k)\right\} \approx \frac{2}{M_j}\sum_{\tau=-\infty}^{\infty} s_{j,k,\tau}^2 \equiv \frac{2A_{j,k}}{M_j}$$

(Keim and Percival, 2012). In practice, we can estimate $A_{j,k}$ using

$$\hat{A}_{j,k} = \frac{1}{2}\left(\hat{\nu}_X^2(\tau_j)\hat{\nu}_X^2(\tau_k) + 2\sum_{\tau=1}^{M_k-1}\hat{s}_{j,\tau}\hat{s}_{k,\tau}\right)$$

(note that, when $k = j$, the above becomes identical to $\hat{A}_j$ of Eq. (23)).

7.1. Estimation of power-law exponents

As illustrated in Section 8 below, plots of the wavelet variance versus scale on log/log axes sometimes show stretches over which $\log(\hat{\nu}_X^2(\tau_j))$ appears to vary linearly with $\log(\tau_j)$, that is, that

$$\log(\hat{\nu}_X^2(\tau_j)) \approx \alpha + \beta \log(\tau_j) \text{ over levels } j, \text{ such that, say, } J_1 \le j \le J_2.$$

This pattern is consistent with a hypothesis that the true wavelet variance obeys a power law over scales τ_{J_1} to τ_{J_2}:

$$\nu_X^2(\tau_j) = c\tau_j^{\beta}, \quad \text{where } c = e^{\alpha}.$$

The power-law exponent β manifests itself as the slope on a log-log plot and is amenable to various interpretations. For example, at small scales, a slope of $0 \le \beta \le 2$ might indicate that $\{X_t\}$ has a fractal dimension of $D = 2 - \frac{\beta}{2}$ (Gneiting et al. (in press)); on the other hand, a log-log plot that is linear over large scales is indicative of an intrinsically stationary process with long-range dependence that might be well modeled by either a fractional Gaussian process with Hurst parameter $H = 1 + \frac{\beta}{2}$ when $-1 < \beta < 0$, a fractional Brownian motion with parameter Hurst parameter $H = \frac{\beta}{2}$ when $0 < \beta < 2$, or a fractionally differenced process with parameter $\delta = (\beta + 1)/2$ when $\beta > -1$ (Abry et al., 1993, 1995; Abry and Veitch, 1998; Coeurjolly, 2008; Faÿ et al., 2009; Flandrin, 1992; Jensen, 1999; Stoev and Taqqu, 2003; Stoev et al., 2006). Metrologists studying fractional frequency deviates from atomic clocks and other high-performance oscillators would equate slopes of $\beta = -3, -2, -1, 0$, and 1 to five canonical noise processes known as white phase, flicker phase, white frequency, flicker frequency, and random-walk frequency noise, respectively (Percival, 2003; Stein, 1985).

To estimate the power-law exponent β based on wavelet variance estimates $\hat{\nu}_X^2(\tau_j)$, $j = J_1, \ldots, J_2$, we first define

$$Y_j \equiv \log\left(\hat{\nu}_X^2(\tau_j)\right) - \psi\left(\tfrac{\eta_j}{2}\right) - \log(2) + \log(\eta_j)$$

and then form the linear regression model

$$Y_j = \alpha + \beta \log(\tau_j) + e_j,$$

for which the error term

$$e_j \equiv \log\left(\frac{\hat{\nu}_X^2(\tau_j)}{\nu_X^2(\tau_j)}\right) - \psi\left(\tfrac{\eta_j}{2}\right) - \log(2) + \log(\eta_j)$$

is equal in distribution to the RV $\log(\chi_{\eta_j}^2) - \psi(\frac{\eta_j}{2}) - \log(2)$. As a rule of thumb, if $\eta_j \ge 10$ for each j, we can regard the e_js as approximately obeying a multivariate Gaussian distribution. In vector notation, we can write

$$\mathbf{Y} = \mathcal{A}\boldsymbol{\theta} + \boldsymbol{e},$$

where $\mathbf{Y} \equiv [Y_{J_1}, \ldots, Y_{J_2}]^T$, $\mathcal{A}$ is a $(J_2 - J_1 + 1) \times 2$ matrix whose first column consists just of ones and whose second column is $\log(\tau_{J_1}), \ldots, \log(\tau_{J_2})$, $\boldsymbol{\theta} \equiv [\alpha, \beta]^T$, and $\boldsymbol{e} \equiv [e_{J_1}, \ldots, e_{J_2}]^T$ is a random vector with zero means and a covariance matrix $\Sigma_{\mathbf{e}}$ that is symmetric with its (j,k)th element given by Eq. (32) when $j \le k$. The generalized least squares (GLS) estimator of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = [\hat{\alpha}, \hat{\beta}]^T \equiv \left(\mathcal{A}^T \Sigma_{\mathbf{e}}^{-1} \mathcal{A}\right)^{-1} \mathcal{A}^T \Sigma_{\mathbf{e}}^{-1} \mathbf{Y} \tag{33}$$

(Draper and Smith, 1998). Under our working assumptions, the estimator $\hat{\boldsymbol{\theta}}$ is multivariate Gaussian with mean vector $\boldsymbol{\theta}$ and covariance matrix $\left(\mathcal{A}^T \Sigma_{\mathbf{e}}^{-1} \mathcal{A}\right)^{-1}$, whose lower right-hand corner is the variance associated with the GLS estimator $\hat{\beta}$ of the power-law exponent. We can in turn use $\hat{\beta}$ to estimate, for example, the parameter δ for a fractionally differenced process via $\hat{\delta} = (\hat{\beta} + 1)/2$, noting that $\operatorname{var}\{\hat{\delta}\} = \operatorname{var}\{\hat{\beta}\}/4$.

A simpler way to estimate $\boldsymbol{\theta}$ is to use a weighted least squares (WLS) estimator. This estimator takes the same form as $\hat{\boldsymbol{\theta}}$ of Eq. (33), but with $\Sigma_{\mathbf{e}}$ replaced by a diagonal matrix $\Lambda_{\mathbf{e}}$ whose diagonal elements are the same as those in $\Sigma_{\mathbf{e}}$; that is, the jth diagonal element is $\operatorname{var}\{\log(\hat{\nu}_X^2(\tau_j))\}$. Using Eqs (32) and (21), we can approximate this variance by

$$\frac{\operatorname{var}\{\hat{\nu}_X^2(\tau_j)\}}{\nu_X^4(\tau_j)} + \frac{4(\operatorname{var}\{\hat{\nu}_X^2(\tau_j)\})^2}{\nu_X^8(\tau_j)} = \frac{2}{\eta_j} + \frac{16}{\eta_j^2}.$$

The WLS estimator is attractive in that it just depends on the EDOFs η_j and does not make use of the covariances $\operatorname{cov}\{\log(\hat{\nu}_X^2(\tau_j)), \log(\hat{\nu}_X^2(\tau_k))\}$, $j \neq k$. The WLS estimator is suboptimal unless these covariances are close to zero, which becomes a better approximation as the wavelet filter width L_1 increases. Assuming the validity of this approximation, we can take the WLS estimator to be multivariate Gaussian with mean vector $\boldsymbol{\theta}$ and covariance matrix $\left(\mathcal{A}^T \Lambda_{\mathbf{e}}^{-1} \mathcal{A}\right)^{-1}$. (Section 9.5 of Percival and Walden (2000) formulates a WLS estimator with the diagonal elements of $\Lambda_{\mathbf{e}}$ given by $\psi'(\frac{\eta_j}{2})$. The tri-gamma function enters into play because $\operatorname{var}\{e_j\} = \operatorname{var}\{\log(\chi_{\eta_j}^2)\} = \psi'(\frac{\eta_j}{2})$, a result due to Bartlett and Kendall (1946). This approach and the one presented here are essentially the same since $\psi'(\frac{\eta_j}{2}) \approx \frac{2}{\eta_j}$ for large η_j.)

7.2. *Estimation of characteristic scale*

The notion of characteristic scale pervades the physical sciences, but has no commonly accepted single definition (von Storch and Zwiers, 1999). Since the wavelet variance is scale based, it is natural to entertain a definition in terms of peaks in plots of $\nu_X^2(\tau_j)$ versus τ_j (Keim and Percival, 2012). Accordingly, suppose $\{X_t\}$ is an intrinsically stationary process whose wavelet variance is such that $\nu_X^2(\tau_j) \ge \nu_X^2(\tau_{j-1})$ and $\nu_X^2(\tau_j) \ge \nu_X^2(\tau_{j+1})$ for some $j \ge 2$, with strict inequality holding in at least one of the two cases. We define a wavelet-based characteristic scale as the location $\tau_{c,j}$ at which a quadratic fit through the points $(x_k, y_k) \equiv (\log(\tau_k), \log(\nu_X^2(\tau_k)))$, $k = j-1$, j, and $j+1$, is maximized:

$$\tau_{c,j} = 2^{-\beta_1/\beta_2} \tau_j, \quad \text{where} \quad \beta_1 \equiv \frac{y_{j+1} - y_{j-1}}{2} \quad \text{and} \quad \beta_2 \equiv y_{j+1} - 2y_j + y_{j-1}.$$

Note that this definition is based only on properties of the wavelet variance locally around scale τ_j, not on its properties at arbitrarily large scales. Other measures of correlation length in use break down in the face of long-range dependence, which is a large-scale property.

We can form an estimator of $\tau_{c,j}$ in an obvious manner by substituting $\hat{y}_k = \log(\hat{\nu}_k^2)$ for y_k in the above equation, thus yielding estimators $\hat{\beta}_1, \hat{\beta}_2$, and hence $\hat{\tau}_{c,j}$. An approximate 95% CI for $\tau_{c,j}$ is given by

$$\left[2^{-1.96\sigma_{\hat{\kappa}}}\hat{\tau}_{c,j}, 2^{1.96\sigma_{\hat{\kappa}}}\hat{\tau}_{c,j}\right]$$

(Keim and Percival, 2012), which depends on the quantity $\sigma_{\hat{\kappa}}$, whose square $\sigma_{\hat{\kappa}}^2$ can be computed through the following steps. Let Σ be the 3×3 covariance matrix whose elements are dictated by Eq. (32); that is, the (m,n)th element of Σ is obtained by setting (j,k) in (32) to $(j-2+m, j-2+n)$, where $m \le n$ range over the values 1, 2, and 3. Let

$$H = \begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} \\ 1 & -2 & 1 \end{bmatrix}.$$

The 2×2 covariance matrix for $\hat{\beta}_1$ and $\hat{\beta}_2$ is given by the symmetric matrix $H\Sigma H^T$. Using the elements of this matrix, we can form

$$\sigma_{\hat{\kappa}}^2 = \frac{\operatorname{var}\{\hat{\beta}_1\}}{\beta_2^2} + \frac{\beta_1^2 \operatorname{var}\{\hat{\beta}_2\}}{\beta_2^4} + \frac{\operatorname{var}\{\hat{\beta}_1\}\operatorname{var}\{\hat{\beta}_2\} + 2(\operatorname{cov}\{\hat{\beta}_1, \hat{\beta}_2\})^2}{\beta_2^4} + \frac{3\beta_1^2(\operatorname{var}\{\hat{\beta}_2\})^2}{\beta_2^6} - \frac{2\beta_1 \operatorname{cov}\{\hat{\beta}_1, \hat{\beta}_2\}}{\beta_2^3}.$$

In practice, we can estimate $\sigma_{\hat{\kappa}}^2$ in a "plug-in" manner by replacing the elements of Σ with obvious estimators.

8. Examples

Here we present five examples of wavelet variance analysis to illustrate the methodology discussed in previous sections.

8.1. Fractional frequency deviates from an atomic clock

We first consider a time series that is derived from measurements of the difference in time as kept by two hydrogen masers. Phase differences ϕ_t between the two masers (directly related to time differences) were measured once per minute for 4000 min and converted into fractional frequency deviates by a proper scaling of the first differences $\phi_t - \phi_{t-1}$. After multiplication by 10^{12} (merely to facilitate plotting), we obtain the series $\{X_t\}$ shown in Fig. 2a. The plot shows several pairs of large positive and negative spikes, which are due to isolated glitches in the phase measurements ϕ_t. The plot

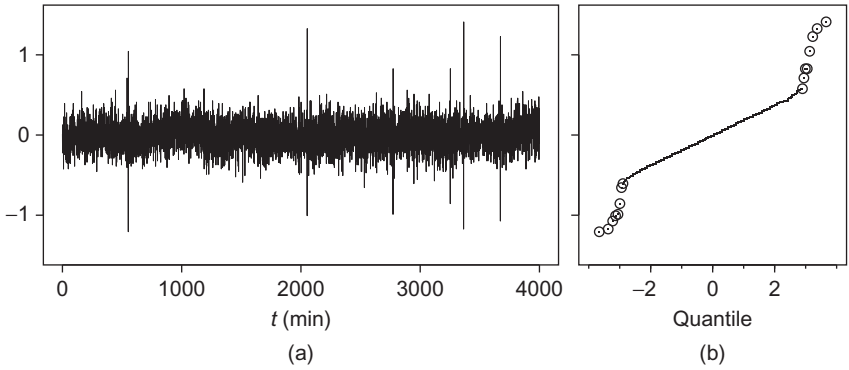

Fig. 2. Fractional frequency deviates (after multiplication by 10^{12}) derived from time differences between two hydrogen masers sampled one per minute, along with a plot of empirical quantiles versus theoretical Gaussian quantiles, with eight smallest and largest quantiles indicated by circles (data courtesy of Drs. Lara Schmidt and Demetrios Matsakis, US Naval Observatory).

of empirical quantiles versus theoretical Gaussian quantiles shown in Fig. 2b demonstrates that the data are well modeled by a Gaussian distribution except for the glitches distorting the tails of the distribution. The glitches are rogue occurrences that do not reflect the inherent ability of the hydrogen masers to keep time.

Scientists assessing the ability of atomic clocks to keep time have used the Allan variance as a performance measure since its introduction in the 1960s (Allan, 1966). The Allan variance is equal to twice the Haar wavelet variance of fractional frequency deviates, so we regard the Haar wavelet and Allan variances as equivalent in the discussion that follows. The popularity of the Allan variance is due in part to the ease with which it can be interpreted relative to SDF-based measures of clock performance (changes in averages over various scales are directly related to timing errors in clocks, whereas the SDF is related only indirectly). Figure 3 shows the Haar wavelet variance estimated using the unbiased estimator $\hat{\nu}_X^2(\tau_j)$ (circles) and the median-type robust estimator $\hat{r}_X^2(\tau_j)$ (diamonds). The conventional and robust estimates are in good agreement, indicating that the rogue values are not adversely affecting $\hat{\nu}_X^2(\tau_j)$.

Plots of the Allan variance have been traditionally used to identify the presence of power-law noise processes, with emphasis on certain canonical laws. Figure 3 makes it clear that no single power-law noise process can adequately model $\{X_t\}$ over all scales, but we can employ different processes over selected scales. For example, arguably the lowest five scales exhibit linear variation in Fig. 3, so we can make use of the methodology described in Section 7.1 to estimate a power-law exponent β over those scales based on $\hat{\nu}_X^2(\tau_j)$ and Eq. (33); however, the exponent so estimated is -1.73, with a corresponding 95% CI of $[-1.76, -1.70]$. This estimate lies between the exponents associated with two canonical power-law processes, namely, flicker-phase noise ($\beta = -2$) and white-frequency noise ($\beta = -1$), but it is not in good agreement with either. Prediction of $\nu_X^2(\tau_j)$ based on the regression model is shown by the line in the lower left-hand portion of Fig. 3, along with asterisks depicting the $\hat{\nu}_X^2(\tau_j)$s (both the line and the wavelet variance estimates are displaced down by an order of magnitude for display purposes).

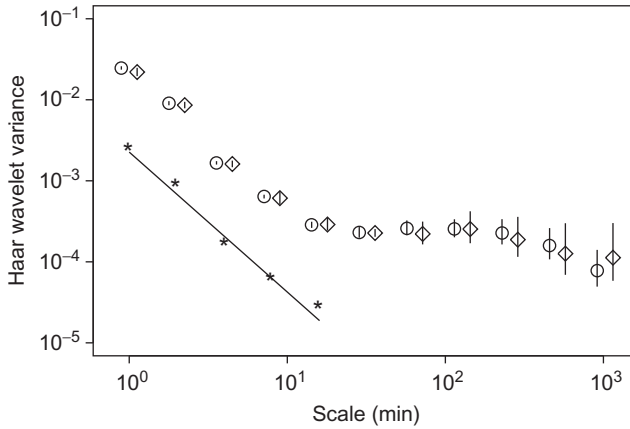

Fig. 3. Haar wavelet variance estimates of fractional frequency deviates along with 95% confidences intervals based on the unbiased estimator $\hat{\nu}_X^2(\tau_j)$ (circles) and median-type robust estimator $\hat{r}_X^2(\tau_j)$ (diamonds). An estimator of the Allan variance (routinely used to assess performance of atomic clocks) is given by $2\hat{\nu}_X^2(\tau_j)$. The asterisks show the unbiased wavelet variance estimates displaced downward by an order of magnitude (i.e., $\hat{\nu}_X^2(\tau_j)/10$) along with a line determined by generalized least squares.

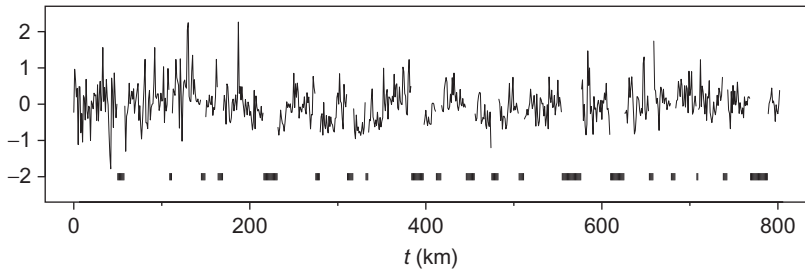

Fig. 4. Residual sea-ice thickness (based on data from a September 1997 Scientific Ice Expedition (SCICEX) cruise archived at the National Snow and Ice Data Center).

8.2. Residual sea-ice thickness

Beginning in the late 1950s, the US Navy used submarines with upward-looking sonars to measure the underwater draft of Arctic sea ice, from which ice thickness can be inferred. The submarines collected these data by cruising under the ice in straight lines, resulting in profiles of thickness along transects that can be treated as a time series (with time being replaced with distance along a transect). These are the most direct observational data we have documenting the evolution of sea-ice thickness over the past half century. Testing the hypothesis that there has been a significant decline in the average thickness of Arctic sea ice requires an understanding of the correlation properties of the profiles. Here we demonstrate how the wavelet variance can be used to assess these properties. Figure 4 shows a thickness profile covering 802 km after detrending by subtracting off a least squares line. This profile has a number of gaps whose positions are indicated by vertical hatch marks under the plot of the profile itself. As described by Percival et al. (2008), we can fill in these gaps using a stochastic interpolation scheme

based upon either a first-order autoregressive or fractionally differenced Gaussian process. Doing so allows us to compute the conventional unbiased Haar wavelet variance estimates $\hat{\nu}_X^2(\tau_j)$ shown by the open circles in Fig. 5a. Rather than using a gap-filled series, we can compute estimates based on the gappy series using the covariance-type estimator of Eq. (28) and the semi-variogram-type estimator of Eq. (29) – these are shown by the solid circles and diamonds, respectively, in Fig. 5a. The three estimates at each scale agree well within one another, which provides some reassurance that the gap-filling procedure is not misrepresenting the correlation properties of the data. Figure 5b replicates the covariance-type estimates. The approximate linear decay of the wavelet variance versus scale on this log–log plot suggests modeling the data as a process with long-range dependence. The WLS estimator described in Section 7.1 gives an estimated power-law exponent of $\hat{\beta} = -0.49$, which translates into an estimate of $\hat{\delta} = 0.26$ for the long-memory parameter for a fractionally differenced process. This

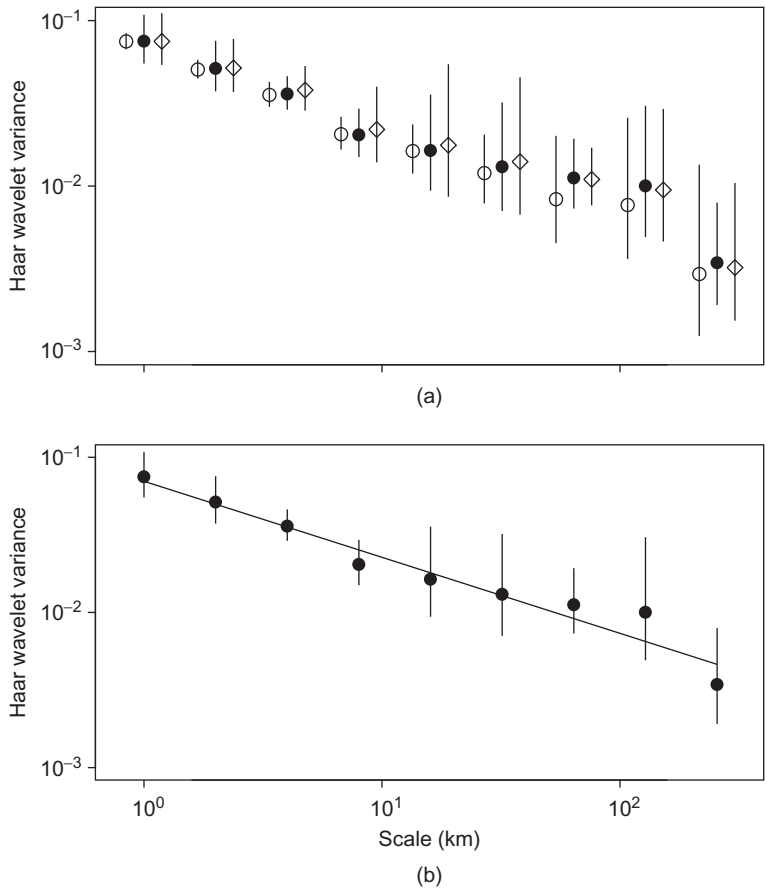

Fig. 5. Haar wavelet variance estimates of residual ice thickness along with 95% confidences intervals based on gap-filled data and the unbiased estimator $\hat{\nu}_X^2(\tau_j)$ (open circles) and on gappy data and the covariance-type and semi-periodogram-type estimators $\hat{u}_X^2(\tau_j)$ and $\hat{v}_X^2(\tau_j)$ (solid circles and diamonds, respectively). The lower plot shows a line from a weighted least squares fit of $\log(\hat{u}_X^2(\tau_j))$ versus $\log(\tau_j)$.

value is consistent with an estimate (0.27) obtained using a computationally intensive maximum likelihood procedure (Percival et al., 2008) and is typical of results obtained for other thickness profiles. The wavelet variance thus contributes to our understanding of the correlation properties of sea-ice thickness, which in turn is a key component in assessing the significance of changes over the past quarter century in this important indicator of the Arctic climate (Rothrock et al., 2008).

8.3. *Albedo measurements of pack ice*

Figure 6 shows a plot of surface albedo (a measure of proportion of incident light reflected) of spring ice in the Beaufort Sea as recorded by the Landsat satellite. The series consists of $N = 8428$ values spaced $\Delta_t = 25$ m apart collected along a transect. Its distribution is highly non-Gaussian, with a lower tail dominated by spikes of low brightness attributable to open water and narrow cracks in thick ice. Lindsay et al. (1996) considered the wavelet variance for this series to investigate its potential for characterizing sea-ice variability. The circles in Fig. 7a show the Haar wavelet variance estimates $\hat{\nu}_X^2(\tau_j)$ for physical scales ranging from 25 m up to 25.6 km (corresponding to standardized scales 1 to 1024). The broad peak in the wavelet variance curve indicates a characteristic scale between 200 and 400 m (standardized scales 8 and 16). The solid curves above and below the circles depict Gaussian-based 95% CIs formed via Eqs (21) through (23), while the dashed curves are corresponding CIs appropriate for non-Gaussian data based on Eqs (25), (27), and then (22) again. Note that, at small scales, the Gaussian-based CIs are considerably narrower than the ones based on non-Gaussian theory, but that the difference is less marked at larger scales, with the two CIs being virtually identical at the largest scale displayed. This example illustrates that there is a danger of underestimating the variability in wavelet variance estimates from an unwarranted assumption of Gaussianity.

Figure 7b again shows the estimates $\hat{\nu}_X^2(\tau_j)$ as circles along with the non-Gaussian-based 95% CIs. The diamonds show the robust median-type estimate $\hat{r}_X^2(\tau_j)$ of Eq. (31) along with 95% CIs. This estimate deemphasizes the spikes in the series and hence reflects background properties of sea ice once the effect of open water and cracks has been downplayed. The robust estimate is quite a bit different from the usual estimate $\hat{\nu}_X^2(\tau_j)$ at small scales, suggesting that these scales are dominated by open water and cracks and that the characteristic scale between 200 and 400 m is mainly due to these

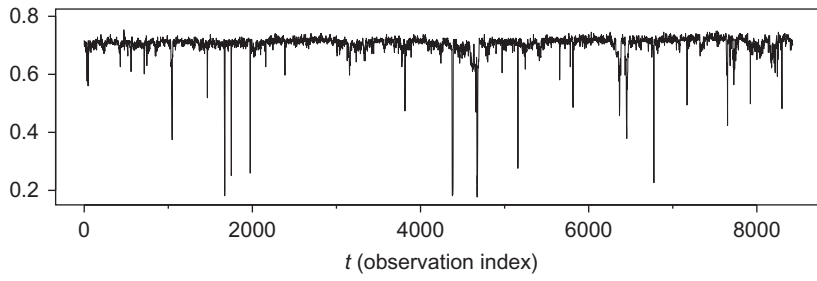

Fig. 6. Surface albedo of pack ice in the Beaufort Sea from a single line of a Landsat TM image obtained from channel 3 on April 16, 1992.

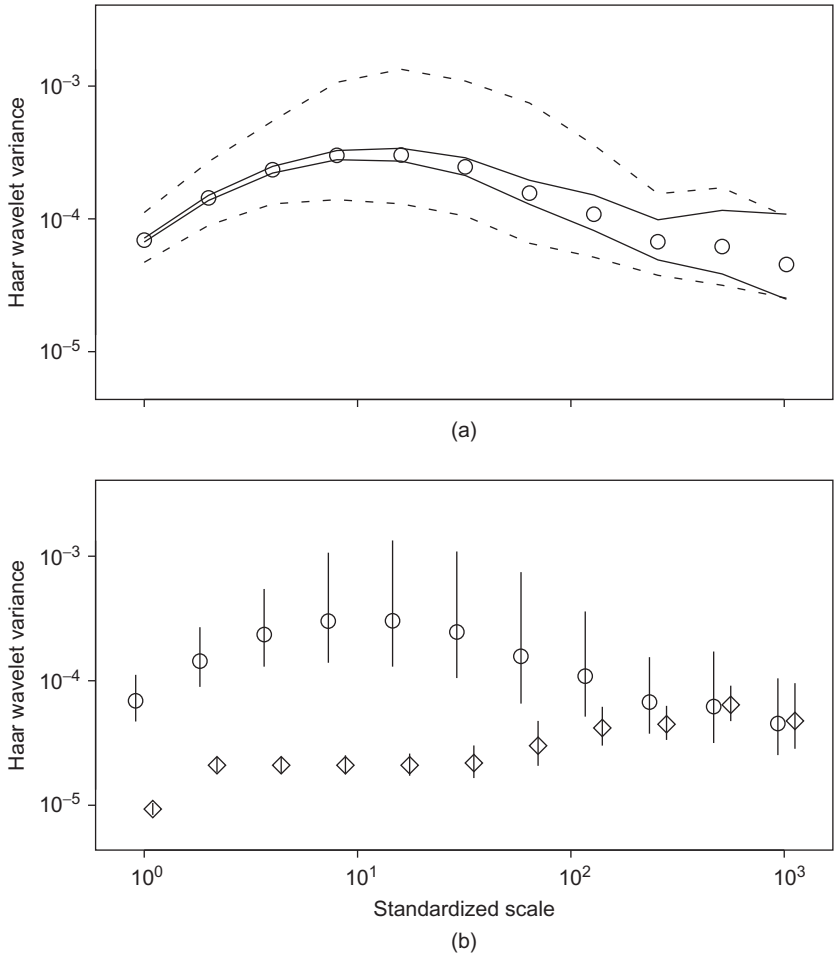

Fig. 7. Haar wavelet variance estimates $\hat{\nu}_X^2(\tau_j)$ for albedo series (circles in both plots), along with Gaussian-based 95% confidence intervals (solid lines in upper plot) and non-Gaussian-based CIs (dashed lines). In addition to $\hat{\nu}_X^2(\tau_j)$ and its associated non-Gaussian-based CIs, the lower plot shows the robust median-type estimate $\hat{r}_X^2(\tau_j)$ (diamonds) with associated 95% CIs.

features. The spatial distribution of these features is of geophysical interest, and hence we cannot regard the spikes in Fig. 6 as rouge observations. Nonetheless, the robust estimate is of interest because it tells us how much of the overall variability at each scale is due to background sea-ice processes that are interrupted by open water and cracks.

8.4. X-ray fluctuations from a binary star system

Our fourth example is a time series of $N = 65,526$ counts from the X-ray binary system GX 5–1 as recorded by the Ginga satellite over a 512-second stretch of time (Hertz and Feigelson, 1997, Norris et al., 1990). Each observation X_t is the number of X-rays

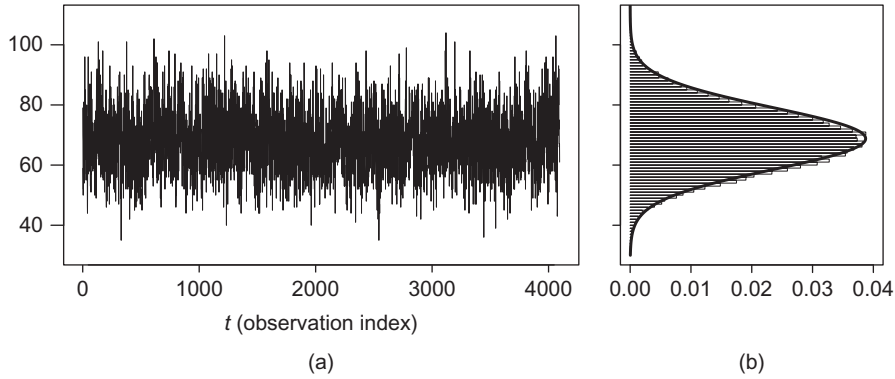

Fig. 8. X-ray fluctuations from a binary star system (first 4096 data values), along with histogram and fitted Gaussian probability density function.

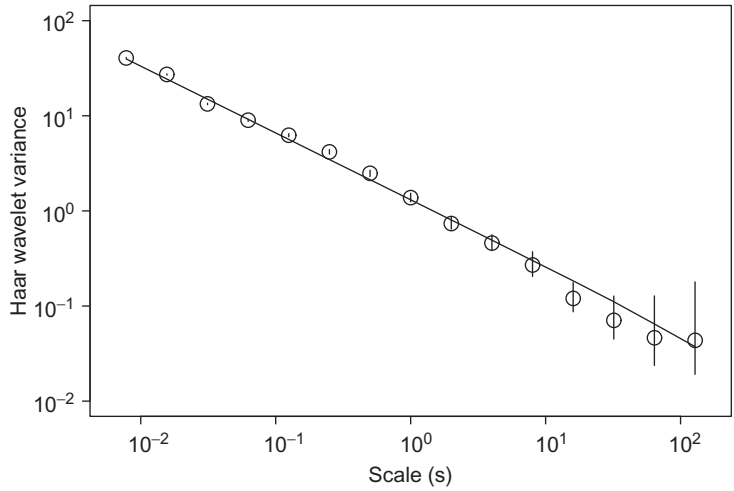

Fig. 9. Unbiased Haar wavelet variance estimates $\hat{\nu}_X^2(\tau_j)$ for the time series of X-ray fluctuations (circles), along with Gaussian-based 95% confidence intervals (vertical lines) and a line with slope -0.7 based upon the generalized least square estimator of Eq. (33).

arriving within an interval (bin size) of $\Delta_t = 1/128$ s. Figure 8a shows the first 4096 values of the time series, while Fig. 8b shows a histogram for the entire series, along with a Gaussian probability density function whose mean and variance match those of the data.

Figure 9 shows a plot of the unbiased Haar wavelet variance estimates $\hat{\nu}_X^2(\tau_j)$ along with 95% CIs, which, since the data are reasonably close to Gaussian, are based on Eqs (21) through (23). Because of the large sample size, the widths of the CIs are so small at the smaller scales as to be barely visible in the plot. Note that $\log(\hat{\nu}_X^2(\tau_j))$ decays roughly linearly with $\log(\tau)$ over all 15 displayed scales. This fact suggests that a power-law model might be a simple description of the overall correlation properties for this time series. The GLS estimator of Eq. (33) yields an estimate of $\hat{\beta} \doteq -0.702$ for the power-law exponent, with an associated 95% CI of $[-0.712, -0.693]$. Using

the relationship $\delta = (\beta + 1)/2$, this exponent translates into an estimate of $\hat{\delta} = 0.149$ for the long-range parameter for a fractionally differenced process. Predicted values for the wavelet variances from the regression model are shown by the line in Fig. 9. Note that the CIs for the wavelet variances fail to trap the predicted values at a number of scales (particularly those below 1 s), an indication that the data have a more intricate correlation structure than what can be captured by a simple power-law model. Thus, the wavelet variance is able both to suggest a simple overall model for the X-ray fluctuations and to point out its limitations.

8.5. *Coherent structures in river flow*

Figure 10 shows a time series capturing so-called coherent structures (such as boils or eddies) in river flows (Chickadel et al., 2009). The 4096 values shown in the plot are from a longer series of length $N = 29,972$ that has a sampling interval of $\Delta = 1/25$ s and spans a little less than 20 min (the subseries in the plot covers about 2.7 min). This time series is derived from measurements from three transducers and a velocity profiler set on the bottom of the Snohomish River Estuary in Washington State immediately downstream of a sill pointing upward. The structures are essentially quasi-periodic upwellings from the river that appear as temporary "blobs" on the surface of the river. Each blob dissipates within a second or so, and then another blob forms sometime later. As the tide increases, the water velocity increases, and the frequency at which the blobs occur appears to increase.

Videos of the river surface clearly show these boils qualitatively, but quantifying this little-understood phenomenon using standard Fourier-based spectral analysis is problematic because it appears as a small perturbation in a low-frequency roll-off. Figure 11 shows unbiased Haar wavelet variance estimates $\hat{\nu}_X^2(\tau_j)$ (circles) for this series, along with associated Gaussian-based 95% CIs (arguably non-Gaussian-based CIs would be more appropriate here). Here, we see a peak at scale $\tau_6\,\Delta = 1.28$ s, a clear indication of a characteristic scale in this vicinity. Using the methodology described in Section 7.2, we obtain an estimated characteristic scale of $\hat{\tau}_{c,6}\,\Delta = 1.6$ s, with an associated 95% CI of [1.4, 1.9] s. The vertical dashed line in Fig. 11 indicates this estimated characteristic scale, whereas the thick horizonal line shows its associated CI. The estimated characteristic scale is based on a quadratic fit through $\hat{\nu}_X^2(\tau_5)$, $\hat{\nu}_X^2(\tau_6)$, and $\hat{\nu}_X^2(\tau_7)$, which is shown by the curve passing through these estimates. In contrast to the SDF, the wavelet

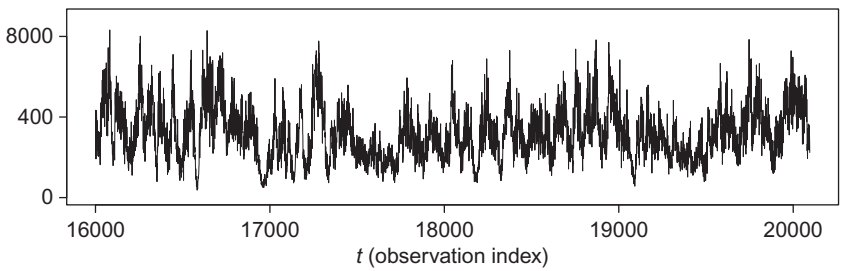

Fig. 10. Coherent structures in river flows (first 4096 values; data courtesy of Alex Horner-Devine and Bronwyn Hayworth, Department of Civil and Environmental Engineering, University of Washington).

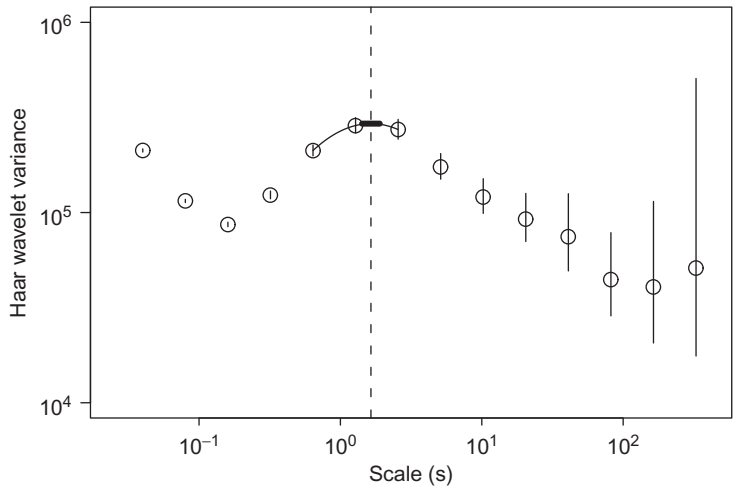

Fig. 11. Haar wavelet variance estimates from time series of coherent structures (circles), with wavelet-based characteristic scale indicated by vertical dashed line. A 95% confidence interval for the characteristic scale is shown by the thick horizontal line, whereas similar intervals for the wavelet variances are shown by the vertical lines emanating from the circles.

variance thus gives a readily interpretable quantification of the phenomenon of interest. We can study the time-evolving properties of the boils by estimating characteristic scales for time series spanning successive 20-min time intervals.

9. Concluding remarks

We have presented a basic introduction to the wavelet variance of time series and its sampling theory. The five real-world examples in the previous section hopefully give the reader an idea of how the wavelet variance can be used as a tool for analyzing time series. Much has been left uncovered. Our statistical treatment has focused on time series that can be regarded as realizations of intrinsically stationary processes. Time series that fall outside of this framework can be fruitfully handled by breaking the series into subseries and analyzing each subseries separately if it is reasonable to assume that each subseries is itself a realization of an intrinsically stationary process (but with a possibly different process for each subseries). This procedure provides a simple way of using the wavelet-based methodology discussed in this chapter to handle certain nonstationary time series; however, the reader should be aware that there are other wavelet-based and -related approaches designed to handle nonstationary time series (see, e.g., Ombao (2012), Chapter 14 in this volume). One approach is based on a "locally stationary modeling" philosophy that facilitates certain asymptotic considerations. A good entry point to this body of literature is Dahlhaus (2012), Chapter 13 in this volume, and references therein.

Our presentation has focused on univariate time series. To study the bivariate relationships between multiple time series, Hudgins (1992) introduced the notion of the wavelet covariance (or wavelet cross spectrum) and wavelet cross-correlation in terms

of a continuous wavelet transform and applied these concepts to atmospheric turbulence in a subsequent paper (Hudgins et al., 1993). Whitcher et al. (2000) and Serroukh and Walden (2000a,b) provide a statistical theory for wavelet covariance analysis of bivariate time series that parallels our treatment of the wavelet variance for single series. This theory presumes that the individual series are intrinsically stationary, but certain bivariate series whose relationships are evolving over time can be handled within this framework by breaking the series into subseries. Sanderson et al. (2010) describe an alternative approach to studying nonstationary bivariate time series that involves the use of wavelet-based locally stationary models. The notion of the wavelet variance can also be extended outside the context of time series to form a scale-based ANOVA for two-dimensional images. Unser (1995) is a pioneering work in this area, which also discusses wavelet-based texture analysis. Lark and Webster (2004) and Milne et al. (2010) document substantive applications of the two-dimensional wavelet variance in the analysis of soil variations. Mondal and Percival (2012b) develop a statistical theory for the two-dimensional wavelet variance that closely parallels the theory for the one-dimensional case presented in this chapter.

Finally, we note that all of the computations and figures in this chapter were done in the statistical language `R` (R Development Core Team, 2011). Code for reproducing all of the numerical examples is available on request from the authors.

Acknowledgments

Preparation of this chapter was supported in part by U.S. National Science Foundation Grant Nos. ARC 0529955 (Percival) and DMS 0906300 (Mondal). Any opinions, findings, and conclusions or recommendations expressed in this chapter are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Abry, P., Gonçalvès, P., Flandrin, P., 1993. Wavelet-based spectral analysis of $1/f$ processes. Proc. IEEE Int. Conf. Acoust. Speech. Signal Process. 3, 237–240.

Abry, P., Gonçalvès, P., Flandrin, P., 1995. Wavelets, spectrum analysis and $1/f$ processes. In: Antoniadis, A., Oppenheim, G. (Eds.), Wavelets and Statistics. Lecture Notes in Statistics, vol. 103. Springer–Verlag, New York, pp. 15–29.

Abry, P., Veitch, D., 1998. Wavelet analysis of long-range-dependent traffic. IEEE Trans. Inform. Theor. 44, 2–15.

Aldrich, E.M., 2005. Alternative Estimators of Wavelet Variance. MS dissertation, Department of Statistics, University of Washington, DC.

Allan, D.W., 1966. Statistics of atomic frequency standards. Proc. IEEE. 54, 221–230.

Bartlett, M.S., Kendall, D.G., 1946. The statistical analysis of variance-heterogeneity and the logarithmic transformation. J. Roy Stat. Soc. Suppl. 8, 128–138.

Beylkin, G., 1992. On the representation of operators in bases of compactly supported wavelets. SIAM J. Numer. Anal. 29, 1716–1740.

Brockwell, P.J., Davis, R.A., 2002. Introduction to Time Series and Forecasting, second ed. Springer, New York.

Bruce, A.G., Gao, H.-Y., 1996. Applied Wavelet Analysis with S-PLUS. Springer, New York.

Chiann, C., Morettin, P.A., 1998. A wavelet analysis for time series. Nonparametric Stat 10, 1–46.

Chickadel, C.C., Horner-Devine, A.R., Talke, S.A., Jessup, A.T., 2009. Vertical boil propagation from a submerged estuarine sill. Geophys. Res. Lett. 36, L10601. doi:10.1029/2009GL037278.

Coeurjolly, J.-F., 2008. Hurst exponent estimation of locally self-similar Gaussian processes using sample quantiles. Ann. Stat. 36, 1404–1434.

Coifman, R.R., Donoho, D.L., 1995. Translation-invariant de-noising. In: Antoniadis, A., Oppenheim, G. (Eds.), Wavelets and Statistics. Lecture Notes in Statistics, vol 103. Springer–Verlag, New York, pp. 125–150.

Craigmile, P.F., Percival, D.B., 2005. Asymptotic decorrelation of between-scale wavelet coefficients. IEEE Trans. Inform. Theor. 51, 1039–1048.

Dahlhaus, R., 2012. Locally Stationary Processes. Elsevier Chapter 13.

Daubechies, I., 1988. Orthonormal bases of compactly supported wavelets. Comm. Pure. Appl. Math. 41, 909–996.

Del Marco, S., Weiss, J., 1997. Improved transient signal detection using a wavepacket-based detector with an extended translation-invariant wavelet transform. IEEE Trans. Signal Process. 45, 841–850.

Draper, N.R., Smith, H., 1998. Applied Regression Analysis, third ed. John Wiley & Sons, New York.

Faÿ, G., Moulines, E., Roueff, F., Taqqu, M., 2009. Estimators of long-memory: Fourier versus wavelets. J. Econometrics. 151, 159–177.

Flandrin, P., 1992. Wavelet analysis and synthesis of fractional Brownian motion. IEEE Trans. Inform. Theor. 38, 910–917.

Giraitis, L., Surgailis, D., 1985. CLT and other limit theorems for functionals of Gaussian processes. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 70, 191–212.

Gneiting, T., Ševčíková, H., Percival, D.B., in press. Estimators of fractal dimension: assessing the roughness of time series and spatial data. Stat. Sci.

Greenhall, C.A., Howe, D.A., Percival, D.B., 1999. Total variance, an estimator of long-term frequency stability. IEEE Trans. Ultrason. Ferroelectrics. Freq. Contr. 46, 1183–1191.

Hertz, P., Feigelson, E.D., 1997. A sample of astronomical time series. In: Subba Rao, T., Priestley, M.B., Lessi, O. (Eds.), Applications of Time Series Analysis in Astronomy and Meteorology. Chapman & Hall, London, pp. 340–356.

Huber, P.J., 1964. Robust estimation of a location parameter. Ann. Math. Stat. 35, 73–101.

Hudgins, L.H., 1992. Wavelet Analysis of Atmospheric Turbulence. PhD dissertation, Department of Physics & Astronomy, University of California, Irvine.

Hudgins, L.H., Friehe, C.A., Mayer, M.E., 1993. Wavelet transforms and atmospheric turbulence. Phys. Rev. Lett. 70, 3279–3282.

Jensen, M.J., 1999. Using wavelets to obtain a consistent ordinary least squares estimator of the long-memory parameter. J. Forecast. 18, 17–32.

Keim, M.J., Percival, D.B., 2012. Assessing characteristic scales using wavelets, Submitted for publication.

Labat, D., Ababou, R., Mangin, A., 2001. Introduction of wavelet analyses to rainfall/runoffs relationship for a karstic basin: the case of Licq–Atherey karstic system. Ground Water 39, 605–615.

Lang, M., Guo, H., Odegard, J.E., Burrus, C.S., Wells, R.O., 1995. Nonlinear processing of a shift invariant DWT for noise reduction. In: Szu, H.H. (Eds.), Wavelet Applications II (Proceedings of the SPIE 2491). SPIE Press, Bellingham, Washington, pp. 640–651.

Lark, R.M., Webster, R., 2001. Changes in variance and correlation of soil properties with scale and location: analysis using an adapted maximal overlap discrete wavelet transform. Eur. J. Soil Sci. 52, 547–562.

Lark, R.M., Webster, R., 2004. Analysing soil variation in two dimensions with the discrete wavelet transform. Eur. J. Soil Sci. 55, 777–797.

Li, T.-H., Oh, H.S., 2002. Wavelet spectrum and its characterization property for random processes. IEEE Trans. Inform. Theor. 48, 2922–2937.

Liang, J., Parks, T.W., 1996. A translation-invariant wavelet representation algorithm with applications. IEEE Trans. Signal Process. 44, 225–232.

Lindsay, R.W., Percival, D.B., Rothrock, D.A., 1996. The discrete wavelet transform and the scale analysis of the surface properties of sea ice. IEEE Trans. Geosci. Rem. Sens. 34, 771–787.

Mallat, S.G., 1989a. Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. Trans. Am. Math. Soc. 315, 69–87.

Mallat, S.G., 1989b. A theory for multiresolution signal decomposition: the wavelet representation IEEE Trans. Pattern Anal. Mach. Intell. 11, 674–693.

Mallat, S.G., 1989c. Multifrequency channel decompositions of images and wavelet models. IEEE Trans. Acoust. Speech Signal Process. 37, 2091–2110.

Mallows, C.L., 1967. Linear processes are nearly Gaussian. J. Appl. Probab. 4, 313–329.

Massel, S.R., 2001. Wavelet analysis for processing of ocean surface wave records. Ocean Eng. 28, 957–987.
Milne, A.E., Lark, R.M., Webster, R., 2010. Spectral and wavelet analysis of gilgai patterns from air photography. Aust. J. Soil Res. 48, 309–325.
Mondal, D., 2007. Wavelet Variance Analysis for Time Series and Random Fields. PhD dissertation, Department of Statistics, University of Washington, DC.
Mondal, D., Percival, D.B., 2010. Wavelet variance analysis for gappy time series. Ann. Inst. Stat. Math. 62, 943–966.
Mondal, D., Percival, D.B., 2012a. M-estimation of wavelet variance analysis. Ann. Inst. Stat. Math. 64, 27–53.
Mondal, D., Percival, D.B., 2012b. Wavelet variance analysis for random fields on a regular lattice. IEEE Trans. Image Process. 21, 537–549.
Nason, G.P., Silverman, B.W., 1995. The stationary wavelet transform and some statistical applications. In: Antoniadis, A., Oppenheim, G. (Eds.), Wavelets and Statistics. Lecture Notes in Statistics, vol. 103. Springer–Verlag, New York, pp. 281–299.
Nason, G.P., von Sachs, R., Kroisandt, G., 2000. Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. J. Roy. Stat. Soc. B. 62, 271–292.
Norris, J.P., Hertz, P., Wood, K.S., Vaughan, B.A., Michelson, P.F., Mitsuda, K., et al., 1990. Independence of short time scale fluctuations of quasi-periodic oscillations and low frequency noise in GX 5–1. Astrophys. J. 361, 514–526.
Ombao, H., 2012. Analysis of Multivariate Non-Stationary Time Series Using the Localized Fourier Library. Elsevier. Chapter 14.
Pelgrum, H., Schmugge, T., Rango, A., Ritchie, J., Kustas, B., 2000. Length-scale analysis of surface albedo, temperature, and normalized difference vegetation index in desert grassland. Water Resour. Res. 36, 1757–1766.
Percival, D.B., 1983. The Statistics of Long Memory Processes. PhD dissertation, Department of Statistics, University of Washington, DC.
Percival, D.B., 1995. On estimation of the wavelet variance. Biometrika 82, 619–631.
Percival, D.B., 2003. Stochastic models and statistical analysis for clock noise. Metrologia 40, S289–S304.
Percival, D.B., Mofjeld, H.O., 1997. Analysis of subtidal coastal sea level fluctuations using wavelets. J. Am. Stat. Assoc. 92, 868–880.
Percival, D.B., Rothrock, D.A., Thorndike, A.S., Gneiting, T., 2008. The variance of mean sea-ice thickness: effect of long-range dependence. J. Geophys. Res. Oceans. 113, C01004. doi:10.1029/2007JC004391.
Percival, D.B., Walden, A.T., 1993. Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques. Cambridge University Press, Cambridge, England.
Percival, D.B., Walden, A.T., 2000. Wavelet Methods for Time Series Analysis. Cambridge University Press, Cambridge, England.
Pesquet, J.-C., Krim, H., Carfantan, H., 1996. Time-invariant orthonormal wavelet representations. IEEE Trans. Signal Process. 44, 1964–1970.
Pichot, V., Gaspoz, J.M., Molliex, S., Antoniadis, A., Busso, T., Roche, F., et al., 1999. Wavelet transform to quantify heart rate variability and to assess its instantaneous changes. J. Appl. Physiol. 86, 1081–1091.
R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/
Rosenblatt, M., 1985. Stationary Sequences and Random Fields. Birkhäuser, Boston.
Rothrock, D.A., Percival, D.B., Wensnahan, M., 2008. The decline in arctic sea-ice thickness: separating the spatial, annual, and interannual variability in a quarter century of submarine data. J. Geophys. Res. Oceans 113, C05003. doi:10.1029/2007JC004252.
Rybák, J., Dorotovič, I., 2002. Temporal variability of the coronal green-line index (1947–1998). Sol. Phys. 205, 177–187.
Sanderson, J., Fryzlewicz, P., Jones, M.W., 2010. Estimating linear dependence between nonstationary time series using the locally stationary wavelet model. Biometrika 97, 435–446.
Scargle, J.D., Steiman-Cameron, T., Young, K., Donoho, D.L., Crutchfield, J.P., Imamura, J., 1993. The quasi-periodic oscillations and very low frequency noise of Scorpius X–1 as transient chaos: a dripping handrail? Astron. J. 411, L91–L94.
Serroukh, A., Walden, A.T., 2000a. Wavelet scale analysis of bivariate time series I: motivation and estimation. Nonparametric Statistics 13, 1–36.

Serroukh, A., Walden, A.T., 2000b. Wavelet scale analysis of bivariate time series II: statistical properties for linear processes. Nonparametric Statistics 13, 36–56.

Serroukh, A., Walden, A.T., Percival, D.B., 2000. Statistical properties and uses of the wavelet variance estimator for the scale analysis of time series. J. Am. Stat. Assoc. 95, 184–196.

Shensa, M.J., 1992. The discrete wavelet transform: wedding the à trous and Mallat algorithms. IEEE Trans. Signal Process. 40, 2464–2482.

Stein, S.R., 1985. Frequency and time – their measurement and characterization. In: Gerber, E.A., Ballato, A., (Eds.), Precision Frequency Control, vol. 2: Oscillators and Standards. Academic Press, Orlando, pp. 191–232.

Stoev, S., Taqqu, M.S., 2003. Wavelet estimation of the Hurst parameter in stable processes. In: Rangarajan, D., Ding, M. (Eds.), Processes with Long Range Correlations: Theory and Applications. Lecture Notes in Physics, vol. 621. Springer, Berlin, pp. 61–87.

Stoev, S., Taqqu, M.S., Park, C., Michailidis, G., Marron, J.S., 2006. LASS: a tool for the local analysis of self-similarity. Comput. Stat. Data Anal. 50, 2447–2471.

Thomson, D.J., 1982. Spectrum estimation and harmonic analysis. Proc. IEEE. 70, 1055–1096.

Torrence, C., Compo, G.P., 1998. A practical guide to wavelet analysis. Bull. Am. Meteorol. Soc. 79, 61–78.

Tsakiroglou, E., Walden, A.T., 2002. From Blackman–Tukey pilot estimators to wavelet packet estimators: a modern perspective on an old spectrum estimation idea. Signal Process. 82, 1425–1441.

Unser, M., 1995. Texture classification and segmentation using wavelet frames. IEEE Trans. Image Process. 4, 1549–1560.

von Storch, H., Zwiers, F.W., 1999. Statistical Analysis in Climate Research. Cambridge University Press, Cambridge, England.

Whitcher, B.J., Byers, S.D., Guttorp, P., Percival, D.B., 2002. Testing for homogeneity of variance in time series: long memory, wavelets and the Nile River. Water Resour. Res. 38, 1054–1070.

Whitcher, B.J., Guttorp, P., Percival, D.B., 2000. Wavelet analysis of covariance with application to atmospheric time series. J. Geophys. Res. Atmos. 105, 14,941–14,962.

Yaglom, A.M., 1958. Correlation theory of processes with random stationary nth increments. Am. Math. Soc. Trans. (Ser. 2). 8, 87–141.

This page intentionally left blank

Part X: Computational Methods

This page intentionally left blank

Time Series Analysis: Methods and Applications, Vol. 30
ISSN: 0169-7161

DOI: 10.1016/B978-0-444-53858-1.00023-5

Time Series Analysis with R

A. Ian McLeod, Hao Yu and Esam Mahdi

Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ontario, Canada N6A 5B7

Abstract

A brief overview of the R statistical computing and programming environment is given that explains why many time series researchers in both applied and theoretical research may find R useful. The core features of R for basic time series analysis are outlined. Some intermediate level and advanced topics in time series analysis that are supported in R are discussed such as including state-space models, structural change, generalized linear models, threshold models, neural nets, co-integration, GARCH, wavelets, and stochastic differential equations. Numerous examples of beautiful graphs constructed using R for time series are shown. R code for reproducing all the graphs and tables is given on my homepage.

Keywords: cluster and multicore computing, quantitative programming environment, reproducible research, statistical computing, time series graphics.

The purpose of our article is to provide a summary of a selection of some of the high-quality published computational time series research using R. A more complete overview of time series software available in R for time series analysis is available in the CRAN[1] task views.[2] If you are not already an R user, this article may help you in learning about the R phenomenon and motivate you to learn how to use R. Existing R users may find this selective overview of time series software in R of interest. Books and tutorials for learning R are discussed later in this section. An excellent online introduction from the R Development Core Team is available[3] as well as extensive contributed documentation.[4]

[1] Comprehensive R Archive.
[2] http://cran.r-project.org/web/views/
[3] http://cran.r-project.org/manuals.html
[4] http://cran.r-project.org/other-docs.html

In the area of computational time series analysis, especially for advanced algorithms, R has established itself as the choice of many researchers. R is widely used not only by researchers but also in diverse time series applications and in the teaching of time series courses at all levels. Naturally, there are many other software systems such as *Mathematica* (Wolfram Research, 2011), that have interesting and useful additional capabilities, such as symbolic computation (Smith and Field, 2001; Zhang and McLeod, 2006). For most researchers working with time series, R provides an excellent broad platform.

The history of R has been discussed elsewhere (Gentleman and Ihaka, 1996), so before continuing our survey, we will just point out some other key features of this quantitative programming environment (QPE).

R is an open source project, providing a freely available and a high-quality computing environment with thousands of add-on packages. R incorporates many years of previous research in statistical and numerical computing, and so it is built on a solid foundation of core statistical and numerical algorithms. The R programming language is a functional, high-level interactive and scripting language that offers two levels of object-oriented programming. For an experienced R user, using this language to express an algorithm is often easier than using ordinary mathematical notation, and it is more powerful since, unlike mathematical notation, it can be evaluated. In this way, R is an important tool of thought. Novice and casual users of R may interact with it using Microsoft Excel (Heiberger and Neuwirth, 2009) or R Commander (Fox, 2005).

Through the use of Sweave (Leisch, 2002, 2003), R supports high-quality technical typesetting and reproducible research including reproducible applied statistical and econometric analysis (Kleiber and Zeileis, 2008). This article has been prepared using Sweave and R scripts for all computations, including all figures and tables, are available in an online supplement.[5] This supplement also includes a PDF preprint of this article showing all graphs in color.

R supports 64-bit, multicore, parallel and cluster computing (Hoffmann, 2011; Revolution Computing, 2011; Schmidberger et al., 2009). Since R is easily interfaced to other programming languages such as C and Fortran, computationally efficient programs may simply be executed in cluster and grid computing environments using R to manage the rather complex message-passing interface.

There is a vast literature available on R that includes introductory books as well as treatments of specialized topics. General purpose introductions to R are available in many books (Adler, 2009; Braun and Murdoch, 2008; Crawley, 2007; Dalgaard, 2008; Everitt and Hothorn, 2009; Zuur et al., 2009). Advanced aspects of the R programming are treated by Chambers (2008), Gentleman (2009), Spector (2008), and Venables and Ripley (2000). Springer has published more than 30 titles in the *Use R* book series, Chapman & Hall/CRC has many forthcoming titles in *The R Series* and there are many other high-quality books that feature R. Many of these books discuss R packages developed by the author of the book and others provide a survey of R tools useful in some application area. In addition to this flood of high-quality books, the *Journal of Statistical Software* (JSS) publishes refereed papers discussing statistical software. JSS reviews not only the paper but also the quality of the computer code as well and publishes both the paper and code on its website. Many of these papers discuss

[5] http://www.stats.uwo.ca/faculty/aim/tsar.html

R packages. The rigorous review process ensures a high-quality standard. In this article, our focus will be on R packages that are accompanied by published books and/or papers in JSS.

The specialized refereed journal, *The R Journal*, features articles of interest to the general R community. There is also an interesting blog sponsored by Revolution Analytics.[6]

The nonprofit association R metrics (Würtz, 2004) provides R packages for teaching and research in quantitative finance and time series analysis that are further described in the electronic books that they publish.

There are numerous textbooks, suitable for a variety of courses in time series analysis (Chan, 2010; Cryer and Chan, 2008; Lütkepohl and Krätzig, 2004; Shumway and Stoffer, 2011; Tsay, 2010; Venables and Ripley, 2002). These textbooks incorporate R usage in the book and an R package on CRAN that includes scripts and datasets used in the book.

1. Time series plots

In this section our focus is on plots of time series. Such plots are often the first step in an exploratory analysis and are usually provided in a final report. R can produce a variety of these plots not only for regular time series but also for more specialized time series such as irregularly spaced time series. The built-in function, `plot()`, may be used to plot simple series such as the annual lynx series, `lynx`. The aspect ratio is often helpful in visualizing slope changes in a time series (Cleveland et al., 1988, Cleveland, 1993). For many time series, an aspect-ratio of 1/4 is good choice. The function `xyplot()` (Sarkar, 2008) allows one to easily control the aspect ratio. Figure 1 shows the time series plot of the lynx series with an aspect ratio of 1/4. The asymmetric rise and fall of the lynx population is easily noticed with this choice of the aspect ratio.

There are many possible styles for your time series plots. Sometimes, a high-density line plot is effective as in Fig. 2.

Another capability of `xyplot()` is the cut-and-stack time series plot for longer series. Figure 3 shows a cut-and-stack plot of the famous Beveridge wheat price index using `xyplot()` and `asTheEconomist()`. The cut-and-stack plot uses the equal count algorithm (Cleveland, 1993) to divide the series into a specified number of subseries using an overlap. The default setting is for a 50% overlap.

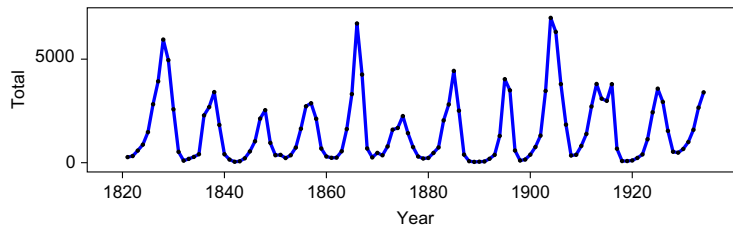

Fig. 1. Annual numbers of lynx trappings in Canada.

[6] http://blog.revolutionanalytics.com/

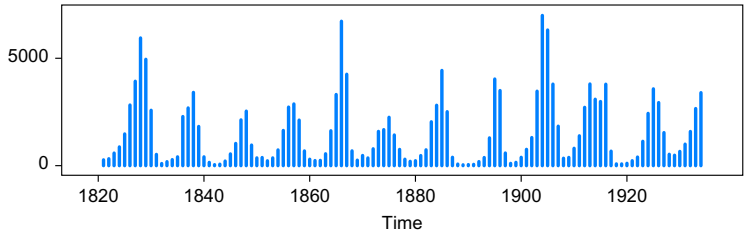

Fig. 2. High-density line plot.

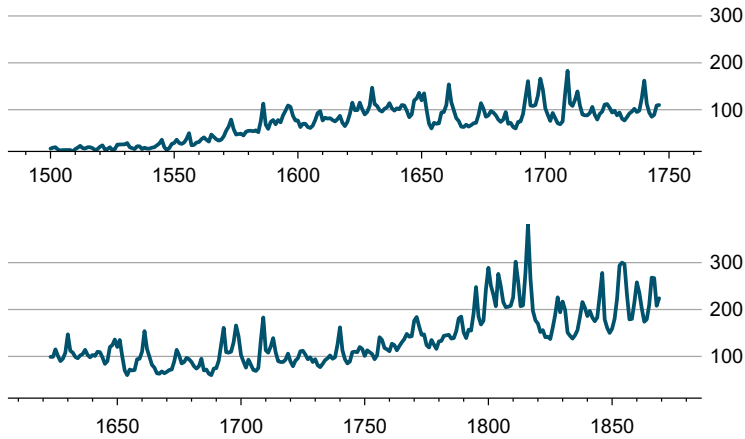

Fig. 3. Beveridge wheat price index.

Figure 4 uses `xyplot()` to plot the seasonal decomposition of the well-known CO_2 time series. The seasonal adjustment algorithm available in R `stl()` is described in the R function documentation and in more detail by Cleveland (1993). This plot efficiently reveals a large amount of information. For example, Fig. 4 reveals that the seasonal amplitudes are increasing.

Bivariate or multivariate time series may also be plotted with `xyplot()`. In Fig. 5, the time series plot for the annual temperature in °C for Canada (CN), Great Britain (UK), and China (CA) 1973–2007, is shown.[7] Figure 5 uses juxtaposition – each series is in a separate panel. This is often preferable to superposition or showing all series in the same panel. Both types of positioning are available using the R functions `plot()` or `xyplot()`.

A specialized plot for bivariate time series called the cave plot (Becker et al., 1994) is easily constructed in R as shown by Zhou and Braun (2010). When there are many multivariate time series, using `xyplot` may not feasible. In this case, `mvtsplot()` provided by Peng (2008) may be used. Many interesting examples, including a stock market portfolio, daily time series of ozone pollution in 100 US counties, and levels of sulfate in 98 US counties are discussed by Peng (2008).

[7] The data were obtained from *Mathematica*'s curated databases.

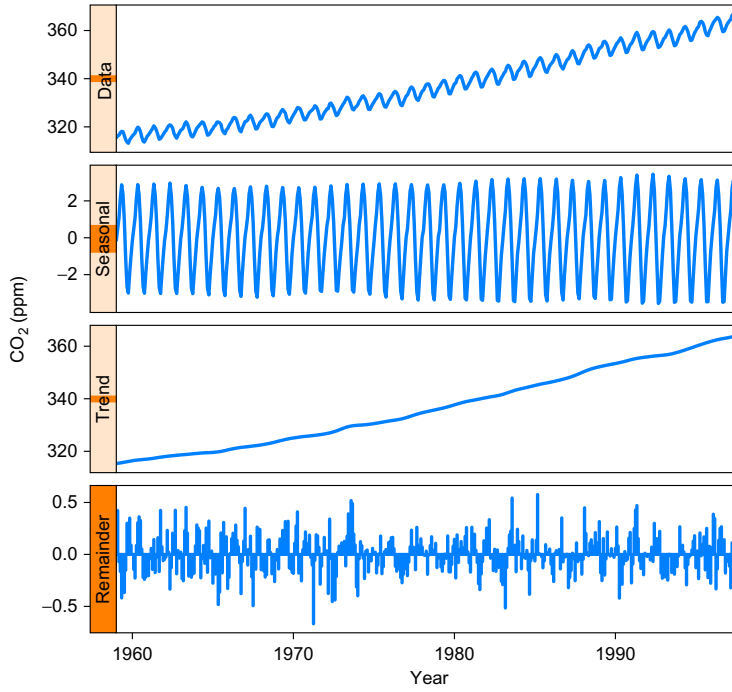

Fig. 4. Atmospheric concentration of CO_2.

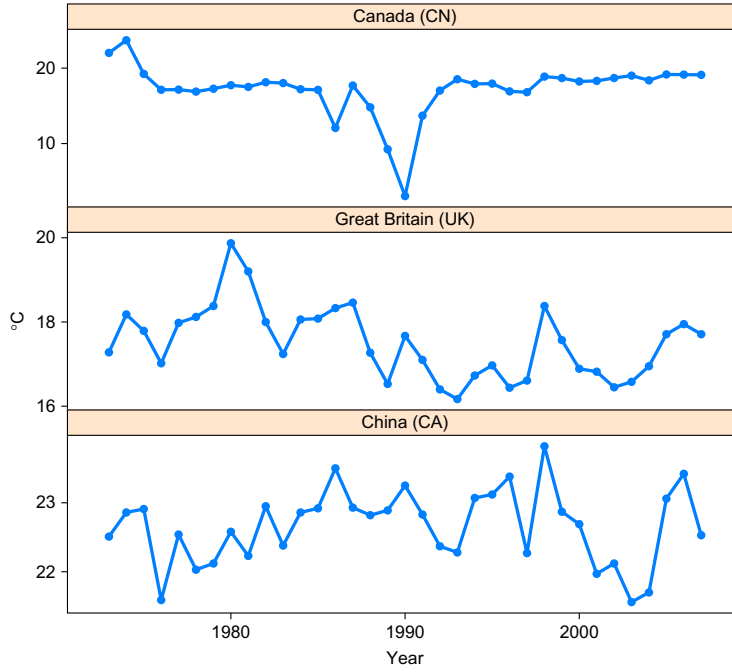

Fig. 5. Average annual temperature (°C) 1973–2007 for Canada (CN), Great Britain (UK), and China (CA).

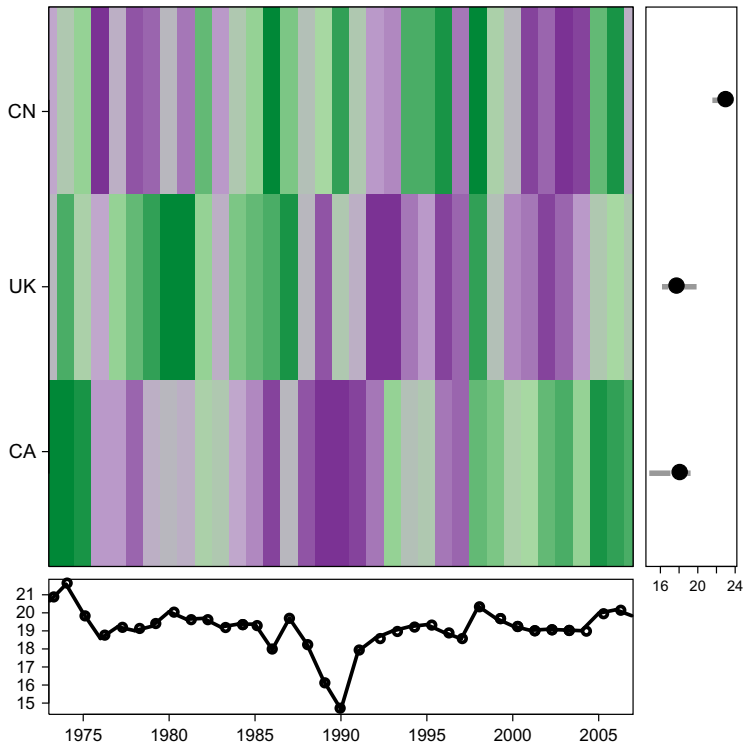

Fig. 6. Average annual temperature in °C, 1973–2007.

Usually, this plot is used with many time series – at least 10 or more – but for simplicity and in order to compare with the last example, Fig. 6 displays the annual temperature series for Canada, Great Britain, and China using using `mvtsplot()`. The right panel of the plot shows a boxplot for the values in each series. From this panel, it is clear that China is generally much warmer than Great Britain and Canada and that Great Britain is often slightly cooler than Canada on an average annual basis. The bottom panel shows the average of the three series. The image shown shows the variation in the three series. The colors purple, grey, and green correspond to low, medium, and high values for each series. The darker the shading, the larger the value. From image in Fig. 6, it is seen that Canada has experienced relatively warmer years than Great Britain or China, since about the year 2000. During 1989–1991, the average annual temperature in Canada was relatively low compared with Great Britain and China. There are many more possible option choices for constructing these plots (Peng, 2008). This plot is most useful for displaying a large number of time series.

Financial time series are often observed on a daily basis but not including holidays and other days when the exchange is closed. Historical and current stock market data may be accessed using `get.hist.quote()` (Trapletti, 2011). Dealing with dates and times is often an important practical issue with financial time series. Grolemund and Wickham (2011) provide a new approach to this problem and review the other approaches that have been used in R. Irregularly observed time series can be plotted

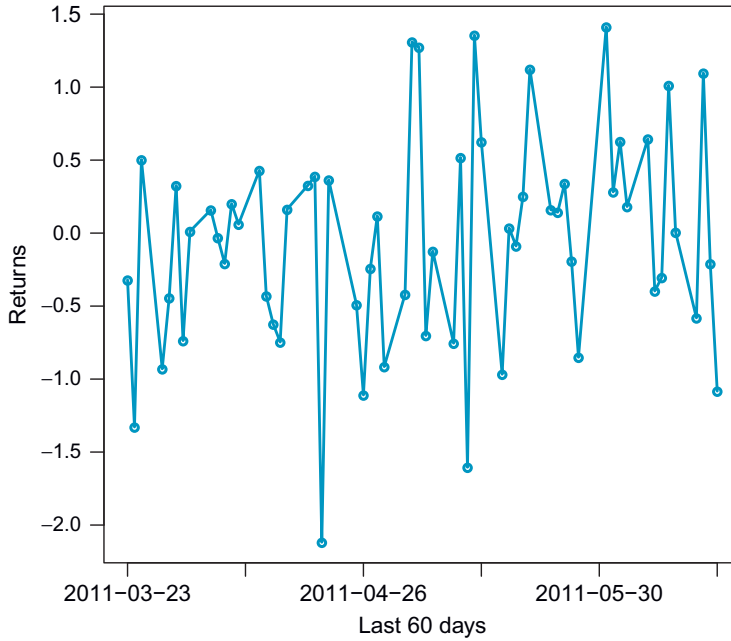

Fig. 7. IBM, daily close price, returns in percent.

using Rmetrics functions (Wuertz and Chalabi, 2011). The RMetrics package **fImport** also has functions for retrieving stock market data from various stock exchanges around the world.

In Fig. 7, the function `yahooSeries()` is used to obtain the last 60 trading days of the close price of IBM stock. The function RMetrics `timeSeries()` converts this data to a format that can be plotted.

Time series plots are ubiquitous and important in time series applications. It must also be noted that R provides excellent time series graphic capabilities with other standard time series functions, including functions time series diagnostics, autocorrelations, spectral analysis, and wavelet decompositions to name a few. The output from such functions is usually best understood from the graphical output.

More generally, there are many other types of functions available for data visualization and statistical graphics. For example, all figures in the celebrated monograph on visualizing data by Cleveland (1993) may be reproduced using the R scripts.[8]

The R package ggplot2 (Wickham, 2009) implements the novel graphical methods discussed in the wonderful graphics book by (Wilkinson, 1999). An interesting rendition of Millard's famous temporal-spatial graph of Napoleon's invasion of Russia using **ggplot2** is available in the online documentation.

Dynamic data visualization, including time series, is provided with **rggobi** (Cook and Swayne, 2007).

[8] http://www.stat.purdue.edu/~wsc/visualizing.html

The foundation and the state-of-the-art in R graphics is presented in the book by Murrell (2011).

2. Base packages: stats and datasets

The **datasets** and **stats** packages are normally automatically loaded by default when R is started. These packages provide a comprehensive suite of functions for analyzing time series, as well as many interesting time series datasets. These datasets are briefly summarized in the Appendix (Section A.1) (Table A.1).

The **stats** package provides the base functions for time series analysis. These functions are listed in the Appendix (A.2) (Tables A.2–A.5). For further discussion of these functions, see the study by Cowpertwait and Metcalfe (2009). Many time series textbooks provide a brief introduction to R and its use for time series analysis (Cryer and Chan, 2008; Shumway and Stoffer, 2011; Venables and Ripley, 2002; Wuertz, 2010).

Adler (2009) provides a comprehensive introduction to R that includes a chapter on time series analysis.

An introduction to ARIMA models and spectral analysis with R is given in the graduate level applied statistics textbook by Venables and Ripley (2002). This textbook is accompanied by the R package **MASS**.

The time series analysis functions that R provides are sufficient to supplement most textbooks on time series analysis.

2.1. stats

First, we discuss the **stats** time series functions. In addition to many functions for manipulating time series such as filtering, differencing, inverse differencing, windowing, simulating, aggregating, and forming multivariate series, there is a complete set of functions for auto/cross correlations analysis, seasonal decomposition using moving-average filters or Loess, univariate and multivariate spectral analysis, univariate and multivariate autoregression, and univariate ARIMA model fitting. Many of these functions implement state-of-the art algorithms. The `ar()` function includes options, in both the univariate and multivariate cases, for Yule-Walker, least-squares or Burg estimates. Although `ar()` implements the maximum likelihood estimator, the package **FitAR** (McLeod et al., 2011b; McLeod and Zhang, 2008b) provides a faster and more reliable algorithm.

The function `spectrum()`, also for both univariate and multivariate series, implements the iterated Daniel smoother (Bloomfield, 2000), and in the univariate case, the autoregressive spectral density estimator (Percival and Walden, 1993).

The `arima()` function implements a Kalman filter algorithm that provides exact maximum likelihood estimation and an exact treatment for the missing values (Ripley, 2002). This function is interfaced to C code to provide maximum computational efficiency. The `arima()` function has options for multiplicative-seasonal ARIMA model fitting, subset models, where some parameters are fixed at zero and regression with ARIMA errors. The functions `tsdiag()` and `Box.test()` provide model diagnostic checks. For ARMA models, a new maximum likelihood algorithm (McLeod and

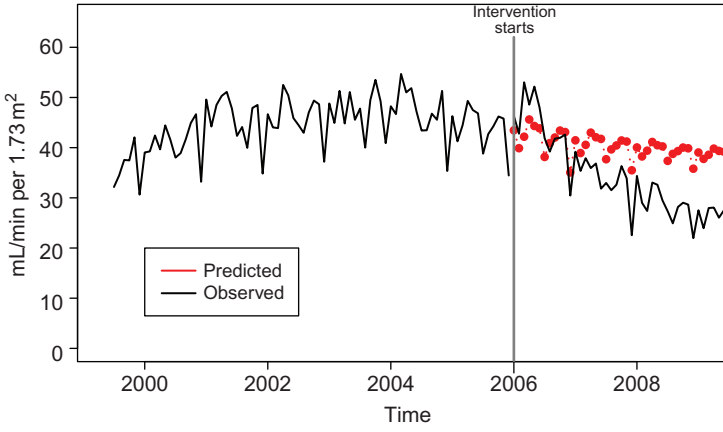

Fig. 8. Creatinine clearance series.

Zhang, 2008a) written entirely in the R language is available in the **FitARMA** package (McLeod, 2010).

A brief example of a medical intervention analysis carried out using `arima()` will now be discussed. In a medical time series of monthly average creatinine clearances, a step intervention analysis model with a multiplicative-seasonal ARIMA$(0,1,1)$ $(1,0,0)_{12}$ error term was fit. The intervention effect was found to be significant at 1%. To illustrate this finding, Fig. 8 compares the forecasts before and after the intervention date. The forecasts are from a model fit to the pre-intervention series. The plot visually confirms the decrease in creatinine clearances after the intervention.

Exponential smoothing methods are widely used for forecasting (Gelper et al., 2010) and are available in **stats** (Meyer, 2002). Simple exponential smoothing defines the prediction for $z_{t+h}, h = 1,2,\ldots$ as $\hat{z}_{t+1}$, where $\hat{z}_{t+1} = \lambda z_t + (1-\lambda)\hat{z}_{t-1}$. The forecast with this method is equivalent to that from an ARIMA(0,1,1). An extension, double exponential smoothing, forecasts $z_{t+h}, h = 1,2,\ldots$ uses the equation $\hat{z}_{t+h} = \hat{a}_t + h\hat{b}_t$, where $\hat{a}_t = \alpha z_t + (1-\alpha)(\hat{a}_{t-1} + \hat{b}_{t-1})$, $\hat{b}_t = \beta(\hat{a}_t - \hat{a}_{t-1}) + (1-\beta)\hat{b}_{t-1}$, where α and β are the smoothing parameters. Double exponential smoothing is sometimes called Holt's linear trend method, and it can be shown to produce forecasts equivalent to the ARIMA(0,2,2). The Winter's method for seasonal time series with period p, forecasts z_{t+h}, by $\hat{z}_{t+h} = \hat{a}_t + h\hat{b}_t + \hat{s}_t$, where $\hat{a}_t = \alpha(z_t - \hat{s}_{t-p}) + (1-\alpha)(\hat{a}_{t-1} + \hat{b}_{t-1})$, $\hat{b}_t = \beta(\hat{a}_t - \hat{a}_{t-1}) + (1-\beta)\hat{b}_{t-1}$, $\hat{s}_t = \gamma(Y - \hat{a}_t) + (1-\gamma)\hat{s}_{t-p}$, α, β, and γ are smoothing parameters. In the multiplicative version, $\hat{z}_{t+h} = (\hat{a}_t + h\hat{b}_t)\hat{s}_t$. Winter's method is equivalent to the multiplicative-seasonal ARIMA airline model in the linear case. All of the above exponential smoothing models may be fit with `HoltWinters()`. This function also has `predict()` and `plot()` methods.

Structural time series models (Harvey, 1989) are also implemented using Kalman filtering in the function `StructTS()`. Since the Kalman filter is used, Kalman smoothing is also available, and it is implemented in the function `tsSmooth()`. The basic structural model is comprised of an observational equation,

$$z_t = \mu_t + s_t + e_t, \quad e_t \sim NID(0, \sigma_e^2)$$

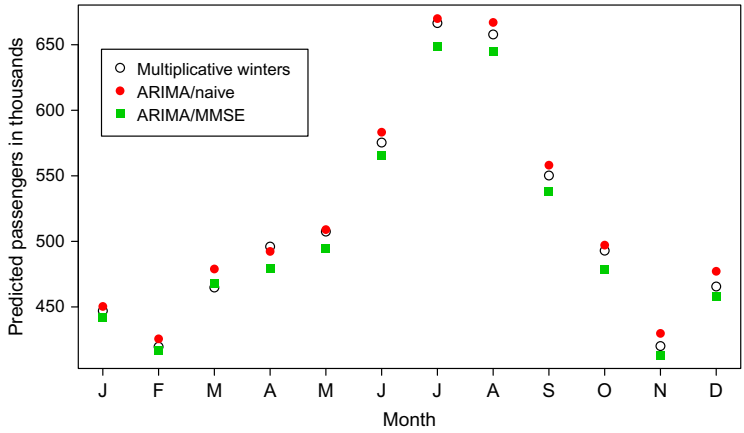

Fig. 9. Comparisons of forecasts for 1961.

and the state equations,

$$\mu_{t+1} = \mu_t + \xi_t, \quad \xi_t \sim \text{NID}(0, \sigma_\xi^2),$$
$$\nu_{t+1} = \nu_t + \zeta_t, \quad \zeta_t \sim \text{NID}(0, \sigma_\zeta^2),$$
$$\gamma_{t+1} = -(\gamma_t + \cdots + \gamma_{t-s+2}) + \omega_t, \quad \omega_t \sim \text{NID}(0, \sigma_\eta^2).$$

If σ_ω^2 is set to zero, the seasonality is deterministic. The local linear trend model is obtained by omitting the term involving γ_t in the observational equation and the last state equation may be dropped as well. Setting $\sigma_\zeta^2 = 0$ in the local linear trend model results in a model equivalent to the ARIMA(0,2,2). Setting $\sigma_\xi^2 = 0$ produces the local linear model, which is also equivalent to the ARMA(0,1,1).

In Fig. 9, the forecasts from the multiplicative Winter's method for the next 12 months are compared with forecasts from the multiplicative-seasonal ARIMA$(0,1,1)$ $(0,1,1)_{12}$ model. With this model, the logarithms of the original data were used, and then the forecasts were backtransformed. There are two types of backtransform that may be used for obtaining the forecasts in the original data domain (Granger and Newbold, 1976; Hopwood et al., 1984) — naive or minimum mean-square error (MMSE). Figure 9 compares these backtransformed forecasts and shows that the MMSE are shrunk relative to the naive forecasts.

2.2. **tseries**

The **tseries** package (Trapletti, 2011) is well established and provides both useful time series functions and datasets. These are summarized in (Appendix A.3).

2.3. **Forecast**

The package **Forecast** (Hyndman, 2010) provides further support for forecasting using ARIMA and a wide class of exponential smoothing models. These methods are

described briefly by Hyndman and Khandakar (2008) and in more depth in the book (Hyndman et al., 2008). Hyndman and Khandakar (2008) discuss a family of 60 different exponential smoothing models and provide a new state-space approach to evaluate the likelihood function.

In Appendix A.4, Table A.10 summarizes functions for exponential smoothing models.

Automatic ARIMA and related functions are summarized in Table A.9.

In addition, general utility functions that are useful for dealing with time series data such as number of days in each month, interpolation for missing values, a new seasonal plot, and others are briefly described in Table A.8.

3. More linear time series analysis

3.1. State-space models and Kalman filtering

Tusell (2011) provides an overview of Kalman filtering with R. In addition to `StructTS`, there are four other packages that support Kalman filtering and state-space modeling of time series. In general, the state-space model (Harvey, 1989; Tusell, 2011) is comprised of two equations, the observation equation:

$$\boldsymbol{y}_t = \boldsymbol{d}_t + \boldsymbol{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t \tag{1}$$

and the state equation:

$$\boldsymbol{\alpha}_t = \boldsymbol{c}_t + \boldsymbol{T}_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{R}_t\boldsymbol{\eta}_t, \tag{2}$$

where the white noises, $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$, are multivariate normal with mean vector zero and covariance matrices Q_t and H_t, respectively. The white noise terms are uncorrelated, $\mathrm{E}\{\boldsymbol{\epsilon}_t'\boldsymbol{\eta}_t\} = 0$.

The Kalman filter algorithm recursively computes

- predictions for $\boldsymbol{\alpha}_t$,
- predictions for $\boldsymbol{y}_t$,
- interpolation for $\boldsymbol{y}_t$,

and in each case, the estimated covariance matrix is also obtained.

Dropping the terms $\boldsymbol{d}_t$ and $\boldsymbol{c}_t$ and restricting all the matrices to be constant over time provides a class of state-space models that includes univariate and multivariate ARMA models (Brockwell and Davis, 1991; Durbin and Koopman, 2001; Gilbert, 1993). As previously mentioned, the built-in function `arima` uses a Kalman filter algorithm to provide exact MLE for univariate ARIMA with missing values (Ripley, 2002). The `dse` package Gilbert (2011) implements Kalman filtering for the time-invariant case and provides a general class of models that includes multivariate ARMA and ARMAX models.

Harrison and West (1997) and Harvey (1989) provide a comprehensive account of Bayesian analysis dynamic linear models based on the Kalman filter, and this theme is further developed in the book by Petris et al. (2009). This book also provides illustrative

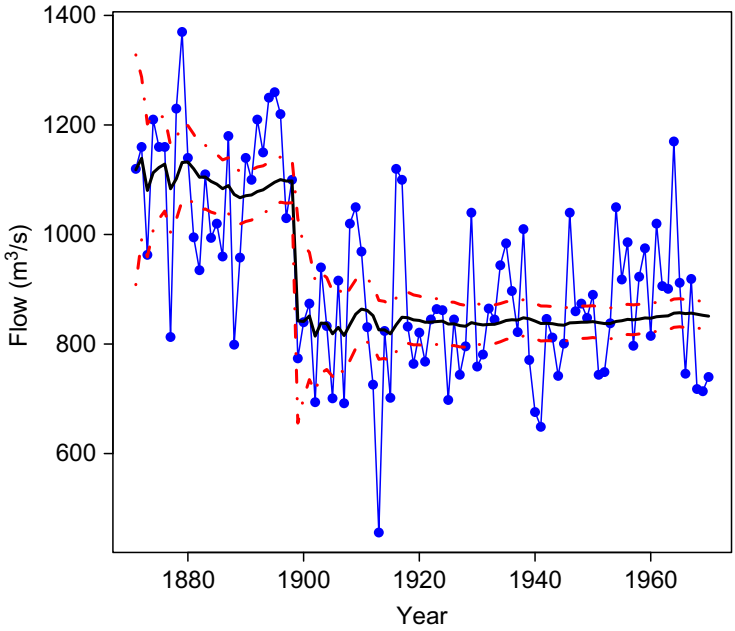

Fig. 10. Nile river flows (solid line with circles), filter values after fitting random walk with noise (solid thick line) and 95% confidence interval (dashed lines).

R scripts and code. The accompanying package **dlm** (Petris, 2010) provides functions for estimation and filtering as well as a well-written vignette explaining how to use the software.

The following example of fitting the random walk plus noise model,

$$y_t = \theta_t + v_t, \quad v_t \sim \mathcal{N}(0, V),$$
$$\theta_t = \theta_{t-1} + w_t, \quad w_t \sim \mathcal{N}(0, W),$$

to the `Nile` series and plotting the filtered series, Fig. 10 and its 95% interval, is taken from the vignette by Petris (2010).

Three other packages for Kalman filtering (Dethlefsen et al., 2009, Luethi et al., 2010, Helske, 2011) are also reviewed by Tusell (2011).

3.2. *An approach to linear time series analysis using Durbin-Levinsion recursions*

Table A.11 in Appendix A.5 lists the main functions available in the **ltsa** package for linear time series analysis.

The Durbin-Levinson recursions (Box et al., 2008) provide a simple and direct approach to the computation of the likelihood, computation of exact forecasts and their covariance matrix, and simulation for any linear process defined by its autocorrelation function. This approach is implemented in **ltsa** (McLeod et al., 2007, 2011a).

In Section 3.3, this approach is implemented for the fractional Gaussian noise (FGN) and a comprehensive model building R package is provided for this purpose using the functions in **ltsa**.

Three methods of simulating a time series given its autocovariance function are available: `DHSimulate()`, `DLSimulate()`, and `SimGLP()`. `DHSimulate()` implements the fast Fourier algorithm (FFT) of Davies and Harte (1987). But this algorithm is not applicable for all stationary series (Craigmile, 2003), so `DHSimulate()`, based on the Durbin-Levinson recursion, is also provided. The algorithm `SimGLP()` is provided for simulating a time series with non-Gaussian innovations based on the equation,

$$z_t = \mu + \sum_{i=1}^{Q} \psi_i a_{t-i}. \tag{3}$$

The sum involved in Eq. (3) is efficiently evaluated using the R function `convolve()` that uses the fast Fourier transform (FFT) method. The built-in function `arima.sim()` may also be used in the case of ARIMA models. The functions `TrenchInverse()` and `TrenchInverseUpdate()` are useful in some applications involving Toeplitz covariance matrices. `TrenchForecast()` provides exact forecasts and their covariance matrix.

The following illustration is often useful in time series lectures when forecasting is discussed. In the next example, we fit an AR(9) to the annual sunspot numbers, 1700–1988, `sunspot.year`. For forecasting computations, it is a standard practice to treat the parameters as known, that is to ignore the error due to estimation. This is reasonable because the estimation error is small in comparison to the innovations. This assumption is made in our algorithm `TrenchForecast()`. Letting $z_m(\ell)$ denote the optimal minimum mean-square error forecast at origin time $t = m$ and lead time ℓ, we compare the forecasts of $z_{m+1}, \ldots, z_n$ using the one-step ahead predictor $z_{m+\ell-1}(1)$, with the fixed origin prediction $z_m(\ell)$, where $\ell = 1, \ldots, L$ and $L = n - m + 1$. Figure 11 compares forecasts and we see many interesting features. The fixed origin forecasts are less accurate as might be expected. As well the fixed origin forecasts show systematic departures, whereas the one step do not.

As shown by this example, `TrenchForecast()` provides a more flexible approach to forecasting than provided by `predict()`.

3.3. Long-memory time series analysis

Let z_t, $t = 1, 2, \ldots$, be stationary with mean zero and autocovariance function, $\gamma_z(k) = \mathrm{cov}(z_t, z_{t-k})$. Many long-memory processes such as the FGN (fractional Gaussian Noise) and FARMA (fractional ARMA) may be characterized by the property that $k^\alpha \gamma_Z(k) \to c_{\alpha,\gamma}$ as $k \to \infty$, for some $\alpha \in (0, 1)$ and $c_{\alpha,\gamma} > 0$. Equivalently,

$$\gamma_Z(k) \sim c_{\alpha,\gamma}\, k^{-\alpha}.$$

The FARMA and FGN models are reviewed by Beran (1994), Brockwell and Davis (1991), Hipel and McLeod (1994). FGN can simply be described as a stationary Gaussian time series with covariance function, $\rho_k = \left(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right)/2$, $0 < H < 1$. The FARMA model generalizes the ARIMA model to a family of stationary models with fractional difference parameter d, $d \in (-0.5, 0.5)$. The long-memory parameters H and d may be expressed in terms of α, $H \simeq 1 - \alpha/2$, $H \in (0, 1)$, $H \neq 1/2$ and $d \simeq 1/2 - \alpha/2$, $d \in (-1/2, 1/2)$, $d \neq 0$ (McLeod,

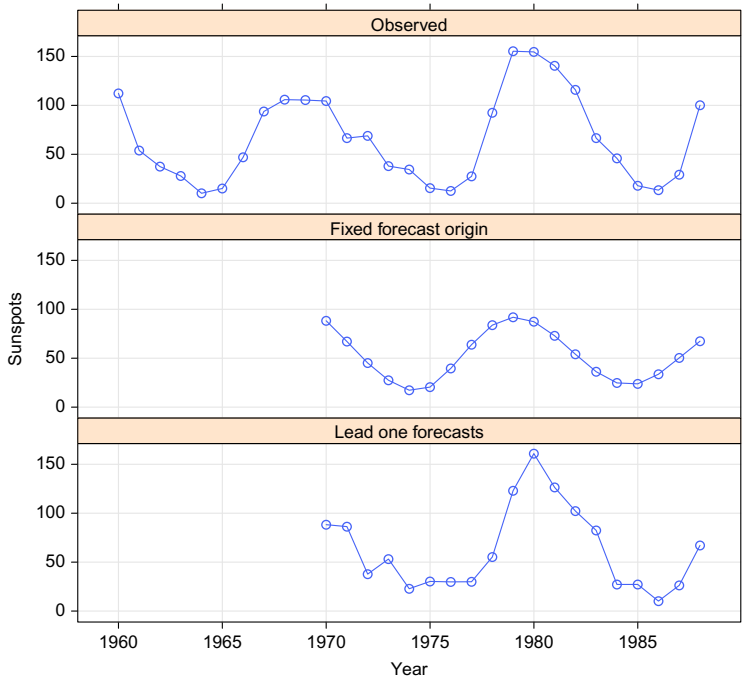

Fig. 11. Comparing forecasts from a fixed origin, 1969, with lead-one forecasts starting in 1969 for `sunspot.year`.

1998). Gaussian white noise corresponds to $H = 1/2$ and in the case of FARMA, $d = 0$ assuming no AR or MA components. Haslett and Raftery (1989) developed an algorithm for maximum likelihood estimation of FARMA models and applied these models to the analysis of long wind speed time series. This algorithm is available in R in the package `fracdiff` (Fraley et al., 2009). The generalization of the FARMA model to allow more general values of d is usually denoted by ARFIMA. A frequently cited example of a long-memory time is the minimum annual flows of the Nile over the period 622–1284, $n = 663$ (Percival and Walden, 2000, Section 9.8). The package `longmemo` (Beran et al., 2009) has this data as well as other time series examples. **FGN** provides exact MLE for the parameter H as well as a parametric bootstrap and minimum mean-square error forecast. For the Nile data, $\hat{H} = 0.831$. The time series plots in Fig. 12 show the actual Nile series along with three bootstraps.

As a further illustration of the capabilities of R, a simulation experiment was done to compare the estimation of the H parameter in fractional Gaussian noise using the exact MLE function `FitFGN()` in **FGN** and the GLM method `FEXPest()` in the package **longmemo**. The function `SimulateFGN()` in **FGN** was used to simulate 100 sequences of length $n = 200$ for $H = 0.3, 0.5, 0.7$. Each sequence was fit by the MLE and GLM method, and the absolute error of the difference between theestimate and the true

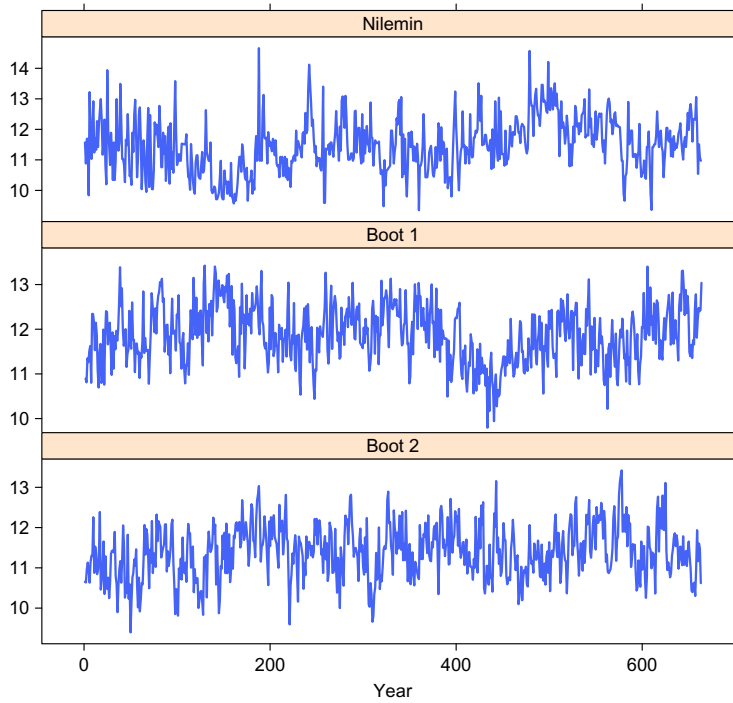

Fig. 12. Comparing actual Nile minima series with two bootstrap versions.

parameter was obtained, that is, $\text{Err}_{\text{MLE}} = |\hat{H}_{\text{MLE}} - H|$ and $\text{Err}_{\text{GLM}} = |\hat{H}_{\text{GLM}} - H|$. From Fig. 13, the notched boxplot for $\text{Err}_{(\text{GLM})} - \text{Err}_{(\text{MLE})}$, we see that the MLE is more accurate. These computations take less than 30 seconds using direct sequential evaluation on a current PC.

The ARFIMA model extends the FARMA models to the ARIMA or difference-stationary case (Baillie, 1996; Diebold and Rudebusch, 1989). The simplest approach is to choose the differencing parameter and then fit the FARMA model to the differenced time series.

3.4. *Subset autoregression*

The **FitAR** package (McLeod and Zhang, 2006, 2008b; McLeod et al., 2011b) provides a more efficient and reliable exact MLE for AR(p) than is available with the built-in function `ar()`. Two types of subset autoregressions may also be fit. The usual subset autoregression may be written, $\phi(B)(z_t - \mu) = a_t$, where $\phi(B) = 1 - \phi_{i_1}B - \cdots - \phi_{i_m}B^{i_m}$, where $i_1, \ldots, i_m$ are the subset of lags. For this model, ordinary least squares (OLS) are used to estimate the parameters. The other subset family is parameterized using the partial autocorrelations as parameters. Efficient model selection, estimation, and diagnostic checking algorithms are discussed by McLeod and Zhang (2006) and McLeod and Zhang (2008b) and implemented in the **FitAR** package (McLeod et al., 2011b). Any stationary time series can be

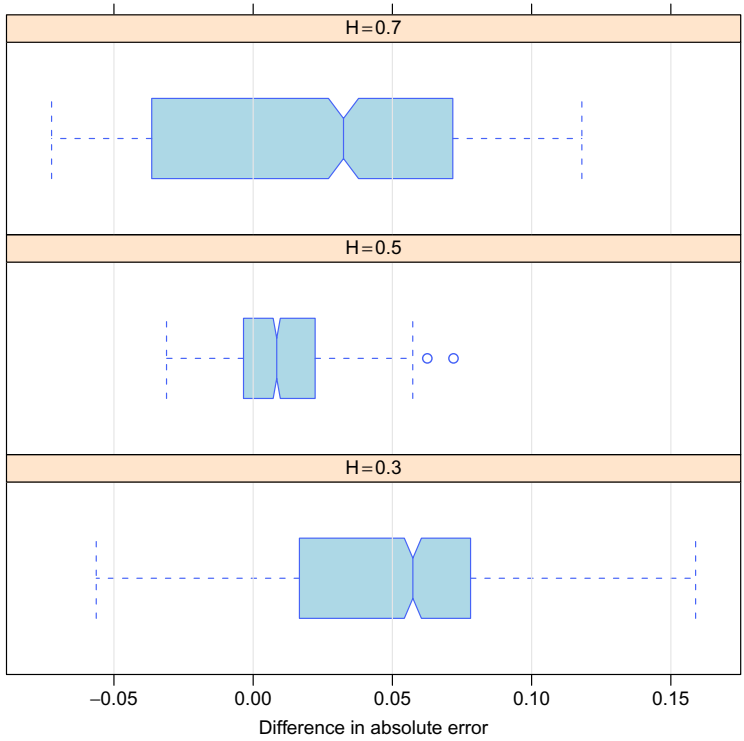

Fig. 13. Comparing MLE estimator and GLM estimator for the parameter H in fractional Gaussian noise.

approximated by a high-order autoregression that may be selected using one of several information criteria. Using this approximation, `FitAR`, provides functions for automatic bootstrapping, spectral density estimation, and Box-Cox analysis for any time series. The optimal Box-Cox transformation for the `lynx` is obtained simply from the command `R > BoxCox(lynx)`. The resulting plot is shown in Fig. 14.

The functions of interest in the **FitAR** package are listed in Appendix A.6.

3.5. *Periodic autoregression*

Let z_t, $t = 1, \ldots, n$ be n consecutive observations of a seasonal time series with seasonal period s. For simplicity of notation, assume that $n/s = N$ is an integer, so N full years of data are available. The time index parameter, t, may be written $t = t(r, m) = (r-1)s + m$, where $r = 1, \ldots, N$ and $m = 1, \ldots, s$. In the case of monthly data, $s = 12$ and r and m denote the year and month. If the expected monthly mean $\mu_m = \mathrm{E}\{z_{t(r,m)}\}$ and the covariance function, $\gamma_{\ell,m} = \mathrm{cov}(z_{t(r,m)}, z_{t(r,m)-\ell})$ depend only on ℓ and m, z_t is said to be periodically autocorrelated and is periodic stationary. The periodic AR model of order $(p_1, \ldots, p_s)$ may be written,

$$z_{t(r,m)} = \mu_m + \sum_{i=1}^{p_m} \phi_{i,m}(z_{t(r,m)-i} - \mu_{m-i}) + a_{t(r,m)},$$

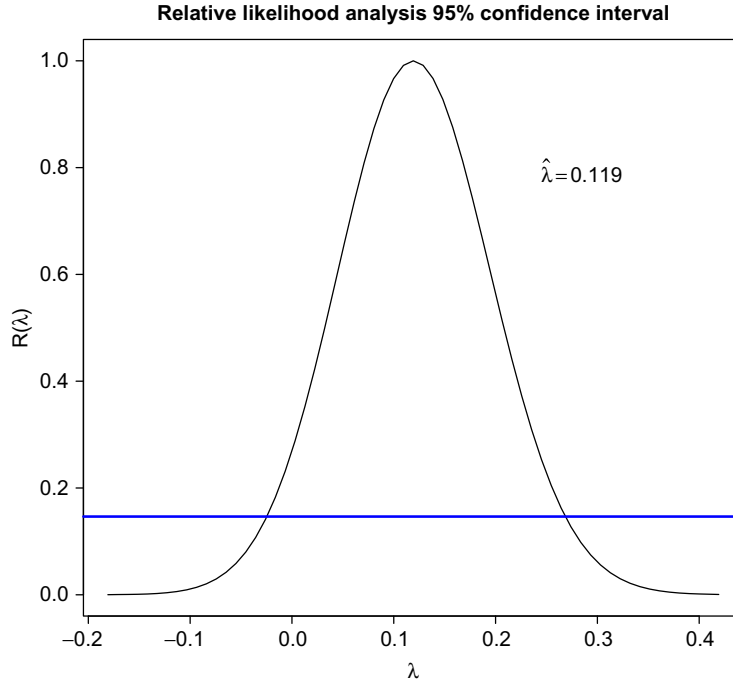

Fig. 14. Box-Cox analysis of lynx time series.

where $a_{t(r,m)} \sim \text{NID}(0, \sigma_m^2)$, where m obeys modular arithmetic base s. This model originated in monthly streamflow simulation and is further discussed with examples by Hipel and McLeod (1994). Diagnostic checks for periodic autoregression are derived by McLeod (1994). The package `pear` (McLeod and Balcilar, 2011) implements functions for model identification, estimation and diagnostic checking for periodic AR models.

We conclude with a brief mention of some recent work on periodically correlated time series models that we hope to see implemented in R. Tesfaye et al. (2011) develop a parsimonious and efficient procedure for dealing with periodically correlated daily ARMA series and provide applications to geophysical series. Ursu and Duchesne (2009) extend modeling procedures to the vector PAR model and provide an application to macroeconomic series. Aknouche and Bibi (2009) show that quasi-MLE provide consistent, asymptotically normal estimates in a periodic GARCH model under mild regularity conditions.

4. Time series regression

An overview of selected time series regression topics is given in this section. Further discussion of these and other topics involving time series regression with R is available in several textbooks (Cowpertwait and Metcalfe, 2009; Cryer and Chan, 2008; Kleiber and Zeileis, 2008; Shumway and Stoffer, 2011).

4.1. Cigarette consumption data

Most of the regression methods discussed in this section will be illustrated with data from an empirical demand analysis for cigarettes in Canada (Thompson and McLeod, 1976). The variables of interest, consumption of cigarettes per capita, Q_t, real disposable income per capita, Y_t, and the real price of cigarettes, P_t, for $t = 1, \ldots, 23$ corresponding to the years 1953–1975 were all logarithmically transformed and converted to an R dataframe `cig`. For some modeling purposes, it is more convenient to use a `ts` object,

```
R >cig.ts <- ts(as.matrix.data.frame(cig), start = 1953,
+     freq = 1)
```

The time series are shown in Fig. 15.

```
R >plot(cig.ts, xlab = "year", main = "", type = "o")
```

4.2. Durbin-Watson test

The exact p value for the Durbin-Watson diagnostic test for lack of autocorrelation in a linear regression with exogenous inputs and Gaussian white noise errors is available with the function `dwtest()` in the **lmtest** package (Hothorn et al., 2010). The diagnostic check statistic may be written

$$d = \frac{\sum_{t=2}^{n}(\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{n}\hat{e}_t^2}, \tag{4}$$

where $\hat{e}_t, t = 1, \ldots, n$ are the OLS residuals. Under the null hypothesis, d should be close to 2 and small values of d indicate positive autocorrelation.

Many econometric textbooks provide tables for the critical values of d. But in small samples, these tables may be inadequate since there is a fairly large interval of values for d for which the test is inconclusive. This does not happen when the exact p value is computed. Additionally, current statistical practice favors reporting p values in diagnostic checks (Moore, 2007).

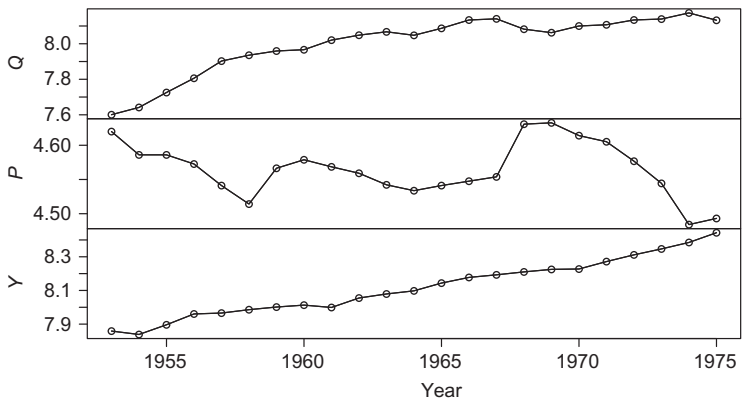

Fig. 15. Canadian cigarette data, consumption/adult (Q), real price (P), income/adult (Y).

The Durbin-Watson test is very useful in time series regression for model selection. When residual autocorrelation is detected, sometimes simply taking first or second differences is all that is needed to remove the effect of autocorrelation. In the next example, we find that taking second differences provides an adequate model.

First, we fit the empirical demand equation, regressing demand Q_t on real price P_t and income Y_t, $Q_t = \beta_0 + \beta_1 P_t + \beta_2 Y_t + e_t$ using OLS with the `lm()` function. Some of the output is shown below.

```
             Estimate Std. Error    t value     Pr(>|t|)
(Intercept)  3.328610  2.5745756  1.2928771 2.107900e-01
P           -0.402811  0.4762785 -0.8457468 4.076991e-01
Y            0.802143  0.1118094  7.1741970 6.011946e-07
```

This output suggests P_t is not significant but Y_t appears to be highly significant. However, since the Durbin-Watson test rejects the null hypothesis of no autocorrelation, these statistical inferences about the coefficients in the regression are incorrect.

After differencing, the Durbin-Watson test still detects significant positive autocorrelation.

Finally, fitting the model with second-order differencing, $\nabla^2 Q_t = \beta_0 + \nabla^2 \beta_1 P_t + \nabla^2 \beta_2 Q_t + e_t$, $\hat{\beta}_1 = 0.557$ with a 95% margin of error, 0.464, so the price elasticity is significant at 5%. As may be seen for the computations reproduced below the other parameters are not statistically significant at 5%.

```
R >cig2.lm <- lm(Q ~ P + Y, data = diff(cig.ts, differences
= 2)) R >summary(cig2.lm)$coefficients
```

```
                Estimate   Std. Error    t value   Pr(>|t|)
(Intercept) -0.003118939  0.008232764 -0.3788447 0.70923480
P           -0.557623890  0.236867207 -2.3541625 0.03012373
Y            0.094773991  0.278979070  0.3397172 0.73800132
```

The intercept term, corresponds to a quadratic trend, is not significant and can be dropped. Income, Y_t, is also not significant. The evidence for lag-one autocorrelation is not strong,

```
R >dwtest(cig2.lm, alternative = "two.sided")
```

```
    Durbin-Watson test

data:  cig2.lm
DW = 2.6941, p-value = 0.08025
alternative hypothesis: true autocorelation is not 0
```

There is also no evidence of non-normality using the Jarque-Bera test. We use the function `jarque.bera.test()` in the **tseries** package (Trapletti, 2011).

```
R >jarque.bera.test(resid(cig2.lm))
```

```
    Jarque Bera Test

data:  resid(cig2.lm)
X-squared = 1.1992, df = 2, p-value = 0.549
```

Kleiber and Zeileis (2008, Section 7) discuss lagged regression models for time series and present illustrative simulation experiment using R that compares the power of the Durbin-Watson test with the Breusch-Godfrey test for detecting residual autocorrelation in time series regression (Kleiber and Zeileis, 2008, Section 7.1).

As discussed below in Section 4.4, fitting regression with lagged inputs is best done using the package **dynlm**.

4.3. Regression with autocorrelated error

The built-in function `arima` can fit the linear regression model with k inputs and ARIMA(p,d,q) errors, $y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + e_t$, where $e_t \sim$ ARIMA(p,d,q) and $t = 1,\ldots,n$.

We illustrate by fitting an alternative to the regression just fit above for the Canadian cigarette data.

```
R >with(cig, arima(Q, order = c(1, 1, 1), xreg = cbind(P,
+     Y)))

Call:
arima(x = Q, order = c(1, 1, 1), xreg = cbind(P, Y))

Coefficients:
         ar1      ma1        P       Y
      0.9332  -0.6084  -0.6718  0.2988
s.e.  0.1010   0.2007   0.2037  0.2377

sigma^2 estimated as 0.0008075:  log likelihood = 46.71,
aic = -83.41
```

This model agrees well with the linear regression using second differencing.

4.4. Regression with lagged variables

Linear regression models with lagged dependent and/or independent variables are easily fit using the **dynlim** package (Zeileis, 2010). In the case of the empirical demand for cigarettes, it is natural to consider the possible effect lagged price. $\nabla^2 Q_t = \beta_1 \nabla^2 P_t + \beta_{1,2} \nabla^2 P_{t-1} + \beta_2 \nabla^2 Y_t + e_t$,

```
R >summary(dynlm(Q ~ -1 + P + L(P) + Y, data = diff(cig.ts,
+     differences = 2)))$coefficients

       Estimate Std. Error    t value   Pr(>|t|)
P    -0.6421079  0.2308323 -2.7817077 0.01278799
L(P) -0.1992065  0.2418089 -0.8238177 0.42145104
Y    -0.2102738  0.2993858 -0.7023507 0.49196623
```

We see that lagged price is not significant.

4.5. Structural change

Brown et al. (1975) introduced recursive residuals and related methods for examining graphically the stability of regression over time. These methods and recent

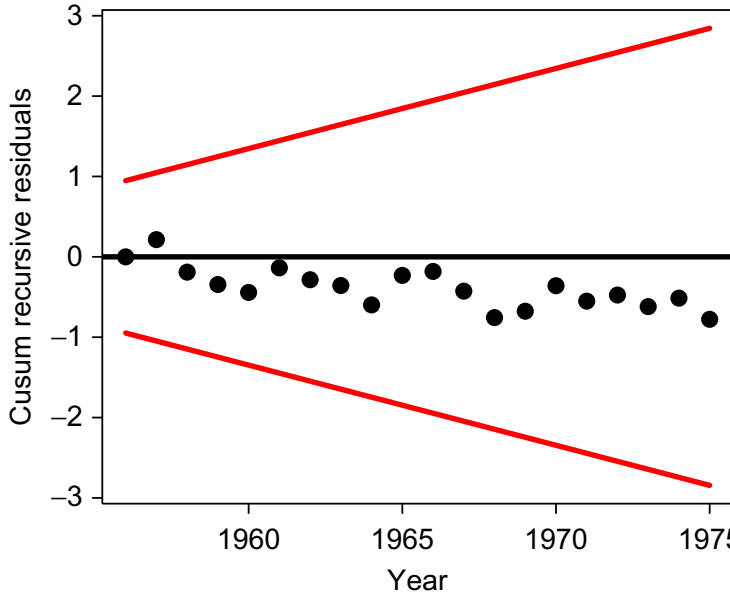

Fig. 16. Cusum test of residuals in cigarette demand regression.

developments in testing and visualizing structural change in time series regression are discussed in the book by Kleiber and Zeileis (2008, Section 6.4) and implemented in the package **strucchange** (Zeileis et al., 2010, 2002). We use a CUMSUM plot of the recursive residuals to check the regression using second differences for stability. As shown in Fig. 16, no instability is detected with this analysis.

4.6. Generalized linear models

Kedem and Fokianos (2002) provide a mathematical treatment of the use of generalized linear models (GLMs) for modeling stationary binary, categorical and count time series. GLMs can account for autocorrelation by using lagged values of the dependent variable in the systematic component. Under regularity conditions, inferences based on large sample theory for GLM time series models can be made using standard software for fitting regular GLMs (Kedem and Fokianos, 2002, Section 1.4). In R, the function `glm()` may be used, and it is easy to verify estimates of the precision using the `boot()` function. These GLM-based time series models are extensively used with longitudinal time series (Li, 1994).

As an illustration, we consider the late night fatality data discussed in Vingilis et al. (2005). The purpose of this analysis was to investigate the effect of the extension of bar closing hours to 2:00 AM that was implemented May 1, 1996. This type of intervention analysis (Box and Tiao, 1975) is known as an interrupted time series design in the social sciences (Shadish et al., 2001). The total fatalities per month for the period starting January 1992 to December 1999, corresponding to a time series of length $n = 84$, are shown in Fig. 17.

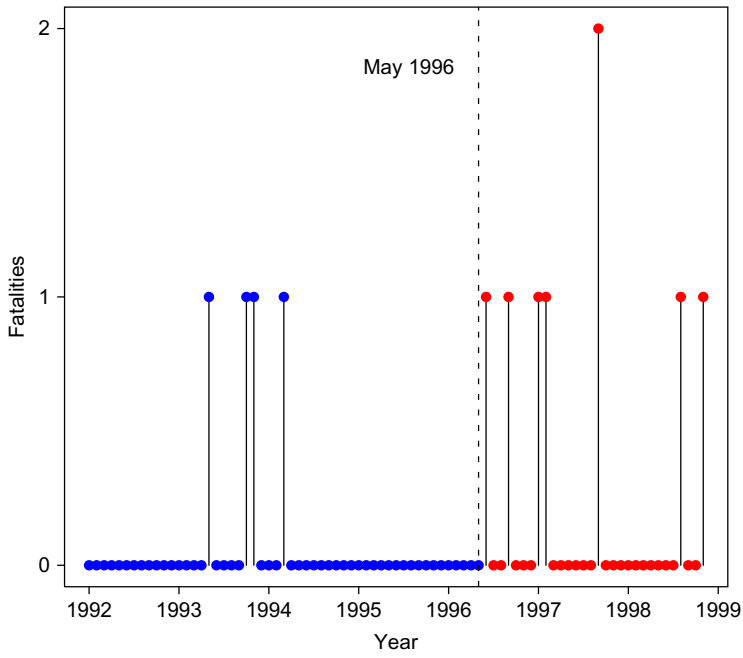

Fig. 17. Late night car fatalities in Ontario. Bar closing hours were extended in May 1996.

The output from the `glm()` function using `y` as the dependent variable, `y1` as the lagged dependent variable,[9] and `x` as the step intervention defined as 0 before May 1, 1996 and 1 after.

```
R >summary(ans)$coefficients

               Estimate Std. Error     z value     Pr(>|z|)
(Intercept) -2.53923499  0.5040873 -5.03729193 4.721644e-07
x2           1.16691417  0.6172375  1.89054329 5.868534e-02
y1          -0.06616152  0.6937560 -0.09536712 9.240232e-01
```

The resulting GLM may be summarized as follows. The total fatalities per month, y_t, are Poisson distributed with mean μ_t, where $\hat{\mu}_t = \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_t + \hat{\beta}_2 y_{t-1}\}$, $\hat{\beta}_0 \doteq -2.54$, $\hat{\beta}_1 \doteq 1.17$, and $\hat{\beta}_2 \doteq -0.07$. There is no evidence of lagged dependence but the intervention effect, β_2, is significant with $p < 0.10$.

We verified the standard deviation estimates of the parameters by using a nonparametric bootstrap with 1000 bootstrap samples. This computation takes less than 10 seconds on most current PC's. Table 1, produced directly from the R output using the package **xtable**, compares the asymptotic and bootstrap standard deviations. As seen from the table the agreement between the two methods is reasonably good.

Hidden Markov models provide another time series generalization of Poisson and binomial GLMs (Zucchini and MacDonald, 2009).

[9] `y` and `y1` are the vectors containing the sequence of observed fatalities and its lagged values.

Table 1
Comparison of asymptotic and bootstrap estimates of the standard deviations in the GLM time series regression

	(Intercept)	$x2$	$y1$
Asymptotic	0.50	0.62	0.69
Bootstrap	0.49	0.66	0.75

5. Nonlinear time series models

Volatility models including the GARCH family of models are one of the newest types on nonlinear time series models. Nonlinear regression models can sometimes be applied to time series. GLMs provide an extension of linear models that is useful for modeling logistic and count time series (Kedem and Fokianos, 2002). Ritz and Streibig (2008) provides an overview of nonlinear regression models using R. Loess regression in R provides a flexible nonparametric regression approach to handling up to three inputs. Using generalized additive models (GAM), many more inputs could be accommodated (Wood, 2006). Two packages, **earth** (Milborrow, 2011) and **mda** (Hastie and Tibshirani, 2011) implement MARS or multiadaptive regression splines (Friedman, 1991). Lewis and Stevens (1991) reported that MARS regression produced better out-of-sample forecasts for the annual sunspot series than competing nonlinear models. In the remainder of the section, we discuss tests for nonlinearity and two popular approaches to modeling and forecasting nonlinear time series, threshold autoregression, and neural net.

5.1. Tests for nonlinear time series

One approach is to fit a suitable ARIMA or other linear time series model and then apply the usual Ljung-Box portmanteau test to the squares of the residuals. McLeod and Li (1983) suggested this as a general test for nonlinearity. The built-in function `Box.test()` provides a convenient function for performing this test. Two tests (Teraesvirta et al., 1993; Lee et al., 1993) for neglected nonlinearity that are based on neural nets are implemented in **tseries** (Trapletti, 2011) as functions `terasvirta.test()` and `white.test()`. The Keenan test for nonlinearity (Keenan, 1985) is available in **TSA** (Chan, 2011) and is discussed in the textbook by Cryer and Chan (2008).

5.2. Threshold models

Threshold autoregression (TAR) provides a general flexible family for nonlinear time series modeling that has proved useful in many applications. This approach is well suited to time series with stochastic cyclic effects such as exhibited in the annual sunspots or lynx time series. The model equation for a two-regime TAR model may

be written,

$$y_t = \phi_{1,0} + \phi_{1,1} y_{t-1} + \cdots + \phi_{1,p} y_{t-p} + I(y_{t-d} > r)\{\phi_{2,0} + \phi_{2,1} y_{t-1} + \cdots + \phi_{2,p} y_{t-p}\} + \sigma a_t \tag{5}$$

where $I(y_{t-d} > r)$ indicates if $y_{t-d} > r$ the result is 1 and otherwise it is 0. The parameter d is the delay parameter and r is the threshold. There are separate autoregression parameters for each regime. This model may be estimated by least squares or more generally using conditional maximum likelihood.

A TAR model for the predator time series in Fig. 18 is described in the book by Cryer and Chan (2008). The package **TSA** (Chan, 2011) provides illustrative datasets from the book (Cryer and Chan, 2008) as well as the function `tar()` for fitting two-regime TAR models, methods functions `predict()` and `tsdiag()`, and functions `tar.skelton()` and `tar.sim()`.

TAR and related models are also discussed by Tsay (2010) and some R scripts are provided as well the companion package **FinTS** (Graves, 2011) that includes datasets from the book. Figure 19 shows monthly US unemployment. Tsay (2010, Example 4.2) fits the two-regime TAR model,

$$\begin{aligned} y_t &= 0.083 y_{t-2} + 0.158 y_{t-3} + 0.0118 y_{t-4} - 0.180 y_{t-12} + a_{1,t} \\ &\quad \text{if } y_{t-1} \le 0.01, \\ &= 0.421 y_{t-2} + 0.239 y_{t-3} - 0.127 y_{t-12} + a_{2,t} \quad \text{if } y_{t-1} > 0.01, \end{aligned}$$

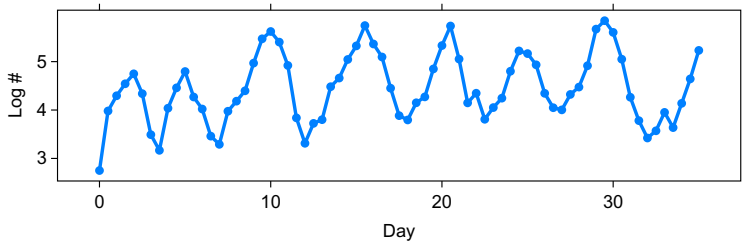

Fig. 18. Number of prey individuals (Didinium natsutum) per ml measured every 12 hours over a period of 35 days.

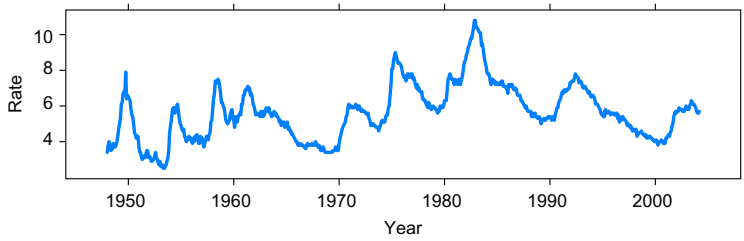

Fig. 19. US civilian unemployment rate, seasonally adjusted, January 1948 to March 2004.

where y_t is the differenced unemployment series. The estimated standard deviations of $a_{2,t}$ and $a_{2,t}$ were 0.180 and 0.217, respectively. Tsay (2010) remarks that the TAR provides more insight into the time-varying dynamics of the unemployment rate than the ARIMA.

5.3. Neural nets

Feed-forward neural networks provide another nonlinear generalization of the autoregression model that has been demonstrated to work well in suitable applications (Faraway and Chatfield, 1998; Hornik and Leisch, 2001; Kajitani et al., 2005). Modeling and forecasting are easily done using **nnet** (Ripley, 2011). A feed-forward neural net that generalizes the linear autoregressive model of order p may be written,

$$y_t = f_o\left(a + \sum_{i=1}^{p} \Omega_i x_i + \sum_{j=1}^{H} w_j f\left(\alpha_j + \sum_{i=1}^{p} \omega_{i,j} x_{t-i}\right)\right), \tag{6}$$

where $\hat{y}_t$ is the predicted time series at time t and $y_{t-1}, \ldots, y_{t-p}$ are the lagged inputs, f_o is the activation function for the output node, f is the activation function for each of the H hidden nodes, $\omega_{i,j}$ are the p weights along the connection for the jth hidden node, Ω_i is the weight in the skip-layer connection, and a is the bias connection. There are $m(1 + H(p + 2))$ unknown parameters that must be estimated. The hyperparameter H, the number of hidden nodes, is determined by a type of cross-validation and is discussed by Faraway and Chatfield (1998), Hornik and Leisch (2001), and Kajitani et al. (2005) in the time series context. The activation functions f and f_o are often chosen to be logistic, $\ell(x) = 1/(1 + e^{-x})$. A schematic illustration for $p = 2$ and $H = 2$ is shown in Fig. 20. Feed-forward neural nets may be generalized for multivariate time series.

Hastie et al. (2009) pointed out that the feed-forward neural net defined in Eq. (6) is mathematically equivalent to the projection pursuit regression model. The net defined in Eq. (6) as well as the one illustrated in Fig. 20 has just one hidden layer with p and $p = 2$ nodes, respectively. These nets may be generalized to accommodate more than one hidden layer and such nets provide additional flexibility. Ripley (1996) shows that asymptotically for a suitable number of hidden nodes, H, and a large enough training sample, the feed-forward neural net with one hidden layer can approximate any continuous mapping between the inputs and outputs.

6. Unit-root tests

Financial and economic time series such as macro/micro series, stock prices, interest rates and many more often exhibit nonstationary wandering behavior. Often, this type of nonstationarity is easily corrected by differencing and the series is said to have a unit root. Such series are sometimes called homogeneous nonstationary or difference stationary. Pretesting for a unit root is useful in ARIMA modeling and in cointegration modeling. Since actual time series may also exhibit other departures from the stationary Gaussian ARMA, many other types of unit-root tests

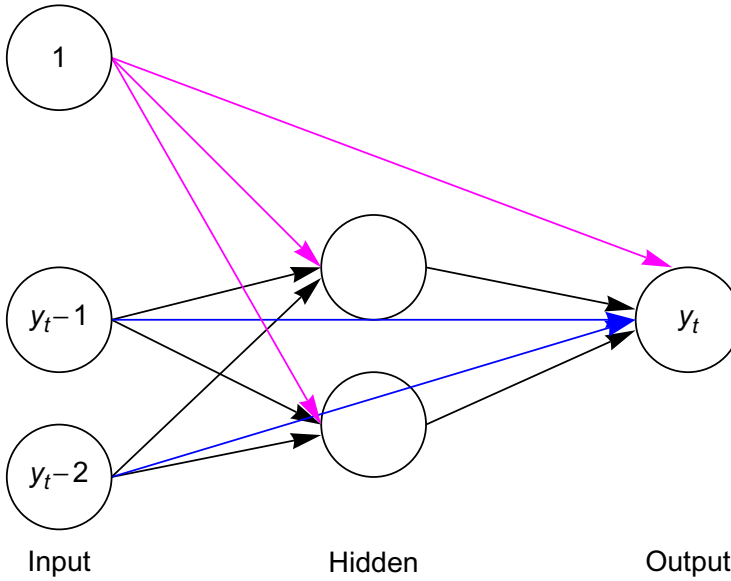

Fig. 20. A nonlinear version of the AR(2) using the feed-forward neural net. This neural net has one hidden layer that is comprised of two hidden nodes. All input nodes have skip-layer connections that connect the input directly with the output.

have been developed that are appropriate under various other assumptions (Elliott et al., 1996; Kwiatkowski et al., 1992; Phillips and Perron, 1988; Said and Dickey, 1984). State-of-the-art testing for unit roots requires a full model building approach that includes taking into account not only possible general autocorrelation effects but also stochastic and deterministic drift components. An incorrect conclusion may be reached if these effects are not taken into account. Such state-of-the-art tests are implemented in the R packages **fUnitRoots** (Wuertz, 2009b) and **urca** (Pfaff, 2010a).

6.1. *Overview of the* **urca** *package*

The **urca** (Pfaff, 2010a) package offers a comprehensive and unified approach to unit-root testing that is fully discussed in the book Pfaff (2006). The textbook by Enders (2010) also provides an excellent overview of the state-of-the-art in unit-root testing. A useful flowchart for using the **urca** package to test for unit roots is given by Pfaff (2006, Chapter 5).

Three regressions with autocorrelated AR(p) errors are considered for the unit-root problem,

$$\Delta Z_t = \beta_0 + \beta_1 t + \gamma Z_{t-1} + \sum_{i=1}^{p-1} \delta_i \Delta Z_{t-i} + e_t \tag{7}$$

$$\Delta Z_t = \beta_0 + \gamma Z_{t-1} + \sum_{i=1}^{p-1} \delta_i \Delta Z_{t-i} + e_t, \tag{8}$$

$$\Delta Z_t = \gamma Z_{t-1} + \sum_{i=1}^{p-1} \delta_i \Delta Z_{t-i} + e_t, \tag{9}$$

corresponding, respectively, to a unit root:

1. with drift term plus deterministic trend,
2. random walk with drift,
3. pure random walk.

The test for unit root corresponds to an upper-tail test of $\mathcal{H}_0 : \gamma = 0$. The parameters β_0 and β_1 correspond to the drift constant and the deterministic time trend, respectively. When $p = 1$, the test reduces to the standard Dickey-Fuller test. To perform the unit-root test, the correct model needs to be identified and the parameters need to be estimated.

The order of the autoregression is estimated using the AIC or BIC. For all three models, the unit-root test is equivalent to testing $\mathcal{H}_0 : \gamma = 0$ is

$$\tau_i = \frac{\hat{\phi} - 1}{\mathrm{SE}(\hat{\phi})}, \quad i = 1, 2, 3,$$

where i denotes the model (9), (8), or (7), respectively. The distribution of τ_i has been obtained by Monte-Carlo simulation or by response surface regression methods (MacKinnon, 1996).

If τ_3 is insignificant, so that $\mathcal{H}_0 : \gamma = 0$ is not rejected, the nonstandard F-statistics Φ_3 and Φ_2 are evaluated using the extra-sum-of-squares principle to test the null hypotheses $\mathcal{H}_0 : (\beta_0, \beta_1, \gamma) = (\beta_0, 0, 0)$ and $\mathcal{H}_0 : (\beta_0, \beta_1, \gamma) = (0, 0, 0)$, respectively. That is, to test whether the deterministic time trend term is needed in the regression model Eq. (7).

If τ_2 is insignificant, so that $\mathcal{H}_0 : \gamma = 0$ is not rejected, the nonstandard F-statistic Φ_1 is evaluated using the extra-sum-of-squares principle to test the hypotheses $\mathcal{H}_0 : (\beta_0, \gamma) = (0, 0)$. That is, to test whether the regression model has a drift term.

If $\mathcal{H}_0 : \gamma = 0$ is not rejected in the final selected model, we conclude that the series has a unit root.

These steps may be repeated after differencing the series to test if further differencing is needed.

6.1.1. Illustrative example

As an example, consider the US real GNP from 1909 to 1970 in billions of US dollars. From Fig. 21, we see there is a strong upward trend. Because the trend does not appear to follow a straight line, a difference-stationary time series model is suggested. This dataset is available as `nporg` in the **urca** package. We set the maximum lag to 4 and use the BIC to select the optimum number of lags. The code snippet is shown below,

```
R >require("urca")
R >data(nporg)
R >gnp <- na.omit(nporg[, "gnp.r"])
R >summary(ur.df(y = gnp, lags = 4, type = "trend",
+     selectlags = "BIC"))
```

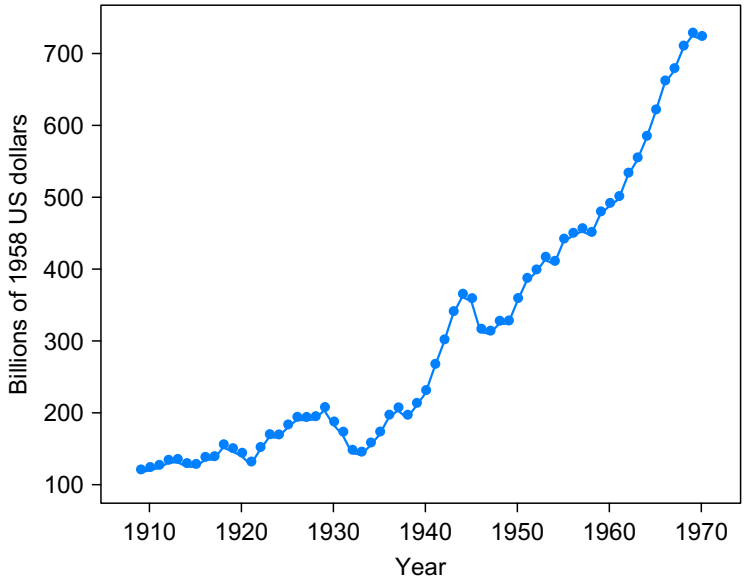

Fig. 21. Real US GNP for 1909–1970.

```
################################################
# Augmented Dickey-Fuller Test Unit Root Test #
################################################


Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-47.149  -9.212   0.819  11.031  23.924

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.89983    4.55369  -0.417  0.67821
z.lag.1     -0.05322    0.03592  -1.481  0.14441
tt           0.74962    0.36373   2.061  0.04423 *
z.diff.lag   0.39082    0.13449   2.906  0.00533 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.19 on 53 degrees of freedom
Multiple R-squared: 0.2727, Adjusted R-squared: 0.2316
F-statistic: 6.625 on 3 and 53 DF,  p-value: 0.0006958
```

```
Value of test statistic is: -1.4814 3.8049 2.7942

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -4.04 -3.45 -3.15
phi2  6.50  4.88  4.16
phi3  8.73  6.49  5.47
```

The above R script fit the full model in Eq. (7) with $p = 4$ and used the BIC to select the final model with $p = 1$. Notice that all test statistics are displayed using the `summary` method.

```
###############################################
# Augmented Dickey-Fuller Test Unit-Root Test #
###############################################

Test regression trend

Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-47.374  -8.963   1.783  10.810  22.794

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.33082    4.02521  -0.082  0.93479
z.lag.1     -0.04319    0.03302  -1.308  0.19623
tt           0.61691    0.31739   1.944  0.05697 .
z.diff.lag   0.39020    0.13173   2.962  0.00448 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
                  0.1 ' ' 1

Residual standard error: 14.88 on 56 degrees of freedom
Multiple R-squared: 0.2684, Adjusted R-squared: 0.2292
F-statistic: 6.847 on 3 and 56 DF,  p-value: 0.0005192


Value of test statistic is: -1.308 3.7538 2.6755

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -4.04 -3.45 -3.15
phi2  6.50  4.88  4.16
phi3  8.73  6.49  5.47
```

When Sweave (Leisch, 2002) is used, Table 2 may be obtained directly from the output produced in R. Figure 22 shows the graphical model diagnostics.

The τ_3 statistic for the null hypothesis $\gamma = 0$ is -1.308, and its corresponding critical values at levels 1%, 5%, and 10% with 62 observations are given in Table 3 as

Table 2
Regression with constant and trend for the US real GNP data starting at 1909 until 1970

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	−0.331	4.025	−0.082	0.935
z.lag.1	−0.043	0.033	−1.308	0.196
tt	0.617	0.317	1.944	0.057
z.diff.lag	0.390	0.132	2.962	0.004

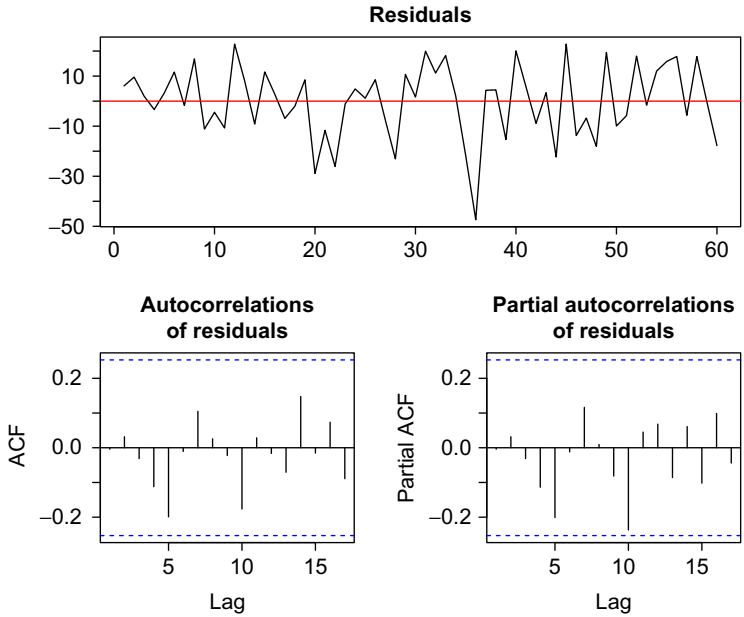

Fig. 22. Residual diagnostic of US real GNP data from 1909 to 1970.

Table 3
Critical values for test statistics for drift and trend case equation (efADFtest1)

	1pct	5pct	10pct
tau3	−4.04	−3.45	−3.15
phi2	6.50	4.88	4.16
phi3	8.73	6.49	5.47

−4.04, −3.45, and −3.15, respectively. At these levels, we can't reject the null hypothesis that $\gamma = 0$ and so we conclude that there is a unit root. Instead of comparing the test statistic value with the critical ones, one can use the MacKinnon's p value

determined from response surface regression methodology (MacKinnon, 1996). The function `punitroot()` is available in **urca**. In the present example, the p value is 0.88, and it corresponds to the τ_3 statistic value confirming that the unit root hypothesis cannot be rejected as in the code snippet below,

```
R >punitroot(result1.ADF@teststat[1], N = length(gnp),
+     trend = "ct", statistic = "t")

[1] 0.8767738
```

The F-statistic Φ_3 is used to test whether the deterministic time trend term is needed in the regression model provided that the model has a drift term. The test statistic has a value of 2.68. From Table 3, the critical values of Φ_3 at levels 1%, 5%, and 10% with 62 observations are 8.73, 6.49, and 5.47, respectively. We conclude that the null hypothesis is not rejected and a trend term is not needed. Thus, we proceed to the next step and estimate the regression parameters in Eq. (8) with a drift term.

```
###############################################
# Augmented Dickey-Fuller Test Unit--Root Test #
###############################################

Test regression drift

Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-47.468  -9.719   0.235  10.587  25.192

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.42944    4.01643   0.356   0.7232
z.lag.1      0.01600    0.01307   1.225   0.2257
z.diff.lag   0.36819    0.13440   2.739   0.0082 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
                0.1 ' ' 1

Residual standard error: 15.24 on 57 degrees of freedom
Multiple R-squared: 0.219,  Adjusted R-squared: 0.1916
F-statistic: 7.993 on 2 and 57 DF,  p-value: 0.0008714


Value of test statistic is: 1.2247 3.5679


Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.51 -2.89 -2.58
phi1  6.70  4.71  3.86
```

Table 4
Regression with drift constant for the US real GNP data

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	1.42944	4.01643	0.35590	0.72323
z.lag.1	0.01600	0.01307	1.22474	0.22571
z.diff.lag	0.36819	0.13440	2.73943	0.00820

Table 5
Dickey-Fuller critical values for test statistics with drift case

	1pct	5pct	10pct
tau2	−3.51	−2.89	−2.58
phi1	6.70	4.71	3.86

Table 6
Critical values for test statistics testing for second differences

	1pct	5pct	10pct
tau3	−4.04	−3.45	−3.15
phi2	6.50	4.88	4.16
phi3	8.73	6.49	5.47

The τ_2 statistic for the null hypothesis $\gamma = 0$ is 1.22474 and its corresponding critical values at levels 1%, 5%, and 10% are given in Table 5 as −3.51, −2.89, and −2.58 respectively. From this analysis we conclude that the series behaves like a random walk with a drift constant term. The next question is whether further differencing might be needed. So we simply repeat the unit root modeling and testing using the differenced series as input.

The τ_3 statistic equals to −4.35. From Table 6, we reject the null hypothesis at 1% and assume that no further differencing is needed.

6.2. *Covariate augmented tests*

The **CADFtest** package (Lupi, 2011) implements Hansen's covariate augmented Dickey-Fuller test (Hansen, 1995) by including stationary covariates in the model equations,

$$a(L)\Delta Z_t = \beta_0 + \beta_1 t + \gamma Z_{t-1} + b(L)'\Delta X_t + e_t \tag{10}$$

$$a(L)\Delta Z_t = \beta_0 + \gamma Z_{t-1} + b(L)'\Delta X_t + e_t, \tag{11}$$

$$a(L)\Delta Z_t = \gamma Z_{t-1} + b(L)'\Delta X_t + e_t. \tag{12}$$

where $a(L) = 1 - a_1 L + \cdots + a_p L^p$ and $b(L)' = b_{q_2} L^{-q_2} + \cdots + b_{q_1} L^{q_1}$. If the main function `CADFtest()` is applied without any stationary covariates, the ordinary

ADF test is performed. In the illustrative example below, taken from the `CADFtest()` online documentation, the augmented test strongly rejects the unit root hypothesis, with a p value less than 2%. On the other hand, with the covariate, the test produces a p value of about 9%. This is shown in the the R session below,

```
R >require(CADFtest)
R >data(npext, package = "urca")
R >npext$unemrate <- exp(npext$unemploy)
R >L <- ts(npext, start = 1860)
R >D <- diff(L)
R >S <- window(ts.intersect(L, D), start = 1909)
R >CADFtest(L.gnpperca ~ D.unemrate, data = S, max.lag.y = 3,
+     kernel = "Parzen", prewhite = FALSE)

    CADF test

data:  L.gnpperca ~ D.unemrate
CADF(3,0,0) = -3.413, rho2 = 0.064, p-value =
0.001729
alternative hypothesis: true delta is less than 0
sample estimates:
      delta
-0.08720302
```

7. Cointegration and VAR models

In the simplest case, two time series that are both difference-stationary are said to be cointegrated when a linear combination of them is stationary. Some classic examples (Engle and Granger, 1987) of bivariate cointegrated series include:

- consumption and income,
- wages and prices,
- short and long-term interest rates.

Further examples are given in most time series textbooks with an emphasis on economic or financial series (Banerjee et al., 1993; Chan, 2010; Enders, 2010; Hamilton, 1994; Lütkepohl, 2005; Tsay, 2010).

A cointegration analysis requires careful use of the methods discussed in these books since spurious relationships can easily be found when working with difference-stationary series (Granger and Newbold, 1974). Most financial and economic time series are not cointegrated. Cointegration implies a deep relationship between the series that is often of theoretical interest in economics. When a cointegrating relationship exists between two series, Granger causality must exist as well (Pfaff, 2006). The **vars** package (Pfaff, 2010b) for vector autoregressive modeling is described in the book (Pfaff, 2006) and article (Pfaff, 2008). This package, along with its companion package **urca** (Pfaff, 2010a), provides state-of-the-art methods for cointegration analysis and modeling stationary and nonstationary multivariate time series.

Full support for modeling, forecasting, and analysis tools are provided for the vector autoregressive time series model (VAR), structural VAR (SVAR), and structural vector

error-correction models (SVEC). The VAR(p) stationary model for a k-dimensional time series, $\{\boldsymbol{y}_t\}$

$$\boldsymbol{y}_t = \boldsymbol{\delta}\boldsymbol{d}_t + \boldsymbol{\Phi}_1\boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{y}_{t-p} + \boldsymbol{e}_t, \tag{13}$$

where $\boldsymbol{\delta}, \boldsymbol{\Phi}_\ell = (\phi_{ij,\ell})_{k\times k}$ are coefficient matrices, $\boldsymbol{d}_t$ is a matrix containing a constant term, linear trend, seasonal indicators or exogenous variables, and $\boldsymbol{\epsilon}_t \sim N(\boldsymbol{0}, I_k)$. Using the **vars** package, the VAR model is estimated using OLS. The basic VAR model, without the covariates $\boldsymbol{d}_t$, may also be estimated using the R core function `ar()`. In the case of the SVAR model,

$$\boldsymbol{A}\boldsymbol{y}_t = \boldsymbol{\delta}\boldsymbol{d}_t + \boldsymbol{\Phi}_1\boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{y}_{t-p} + \boldsymbol{B}\boldsymbol{e}_t, \tag{14}$$

where $\boldsymbol{A}$, and $\boldsymbol{B}$ are $k \times k$ matrices. With the structural models, further restrictions are needed on the parameters and after the model has been uniquely specified, it is estimated by maximum likelihood. The SVEC model is useful for modeling nonstationary multivariate time series and is an essential tool in cointegration analysis. The basic error-correction model, VEC, may be written,

$$\nabla\boldsymbol{y}_t = \boldsymbol{\Pi}\boldsymbol{y}_t + \boldsymbol{\Gamma}_1\nabla\boldsymbol{y}_{t-1} + \cdots + \nabla\boldsymbol{\Gamma}_p\boldsymbol{y}_{t-p+1} + \boldsymbol{e}_t, \tag{15}$$

where ∇ is the first-differencing operator and $\boldsymbol{\Pi}$ and $\boldsymbol{\Gamma}_\ell, \ell = 1, \ldots, p-1$ are parameters. As with the VAR model, the VEC model may be generalized to the SVEC model with coefficient matrices $\boldsymbol{A}$ and/or $\boldsymbol{B}$. A cointegration relationship exists provided that $0 < \text{rank}\ \boldsymbol{\Pi} < p$. When rank $\boldsymbol{\Pi} = 0$, a VAR model with the first differences may be used, and when $\boldsymbol{\Pi}$ is of full rank, a stationary VAR model of order p is appropriate. The **vars** package includes functions for model fitting, model selection, and diagnostic checking as well as forecasting with VAR, SVAR, and SVEC models. Cointegration tests and analysis are provided in the **urca**. In addition to the two-step method of Engle and Granger (1987), tests based on the method of Phillips and Ouliaris (1990) and the likelihood method (Johansen, 1995) are implemented in the **urca** package. Illustrative examples of how to use the software for multivariate modeling and cointegration analysis are discussed in the book, paper, and packages of Pfaff (2006, 2008, 2010b).

8. GARCH time series

Volatility refers to the random and autocorrelated changes in variance exhibited by many financial time series. The GARCH family of models (Engle, 1982; Bollerslev, 1986) capture quite well volatility clustering as well as the thick-tailed distributions often found with financial time series such as stock returns and foreign exchange rates. The GARCH family of models is discussed in more detail in textbooks dealing with financial time series (Chan, 2010; Cryer and Chan, 2008; Enders 2010; Hamilton, 1994; Shumway and Stoffer, 2011; Tsay, 2010).

A GARCH(p, q) sequence $a_t, t = \ldots, -1, 0, 1, \ldots$ is of the form

$$a_t = \sigma_t\epsilon_t$$

and

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_j a_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2,$$

where $\alpha_0 > 0$, $\alpha_i \geq 0$, $1 \leq i \leq p$, $\beta_j \geq 0$, $1 \leq j \leq q$ are parameters. The errors ϵ_t are assumed to be independent and identically distributed from a parametric distribution such as normal, generalized error distribution (GED), Student-t or skewed variations of these distributions. Although ARMA models deal with nonconstant conditional expectation, GARCH models handle nonconstant conditional variance. Sometimes, those two models are combined to form the ARMA/GARCH family of models. A comprehensive account of these models is also given in the book by Zivot and Wang (2006). This book also serves as the documentation for the well-known S-Plus add-on module, **Finmetrics**. Many of the methods provided by **Finmetrics** for GARCH and related models are now available with the **fGARCH** package (Wuertz, 2009a). In the following, we give a brief discussion of the use of **fGARCH** for simulation, fitting, and inferences. The principal functions in this package include `garchSpec`, `garchSim`, and `garchFit` and related methods functions. The **fGarch** package allows for a variety of distributional assumptions for the error sequence ϵ_t. As an illustrative example, we simulate a GARCH(1,1) with $\alpha_0 = 10^{-6}$, $\alpha_1 = 0.2$, and $\beta_1 = 0.7$ and with a skewed GED distribution with skewness coefficient 1.25 and shape parameter 4.8. The simulated series is shown in Fig. 23.

```
R> require("fGarch")
R> spec <- garchSpec(model = list(omega = 1e-06,
```

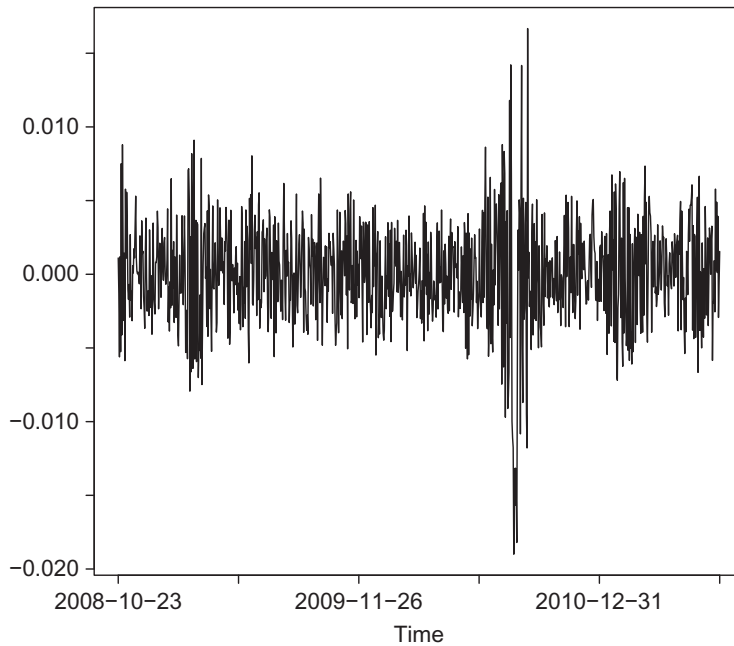

Fig. 23. Simulated GARCH(1, 1) with $\alpha_0 = 10^{-6}, \alpha_1 = 0.2, \beta_1 = 0.7$.

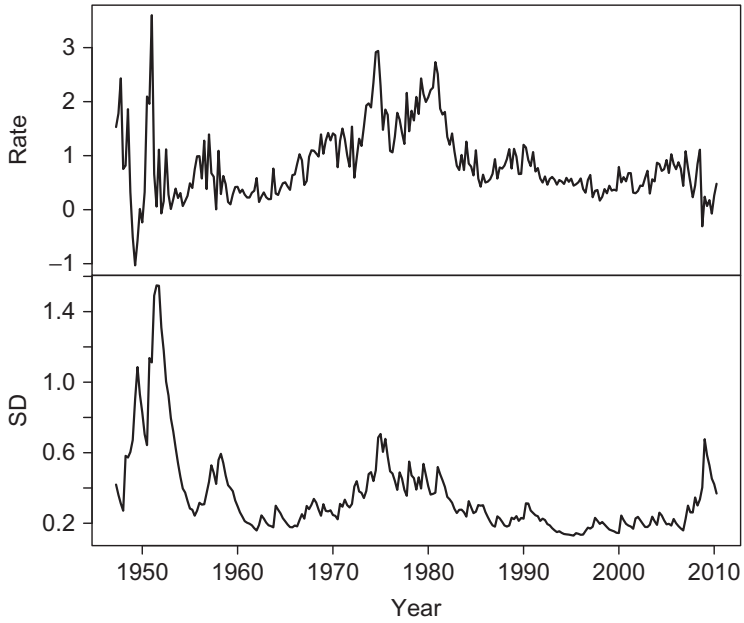

Fig. 24. Inflation rate, r_t, and volatility, σ_t.

```
   alpha = 0.2,
+     beta = 0.7, skew = 1.25, shape = 4.8), cond.dist =
   "sged")
R> x <- garchSim(spec, n = 1000)
```

To fit the above simulated data with GARCH(1,1), we could use,

```
R> out <- garchFit(~garch(1, 1), data = x, trace = FALSE)
```

Some of the inferences that can be carried out by using the `summary()` function, include the Jarque-Bera and Shapiro-Wilk normality tests, various Ljung-Box white noise tests, and ARCH effect tests.

As a further illustration, we fit an ARMA/GARCH model to the US inflation (Bollerslev, 1986). We used the GNP deflator for 1947-01-01 to 2010-04-01. There were $n = 254$ observations that are denoted by $z_t, t = 1, \ldots, n$. Then, the inflation rate may be estimated by the logarithmic difference, $r_t = \log(z_t) - \log(z_{t-1})$. The following ARMA/GARCH model was fit using the function `garchFit()` in **fGarch**, $r_t = 0.103 + 0.369r_{t-1} + 0.223r_{t-2} + 0.248r_{t-3} + \epsilon_t$, and $\sigma_t^2 = 0.004 + 0.269\epsilon_{t-1}^2 + 0.716\sigma_{t-1}^2$. Figure 24 shows time series plots for r_t and σ_t. The **tseries** (Trapletti, 2011) can also fit GARCH models but **fGarch** provides a more comprehensive approach.

9. Wavelet methods in time series analysis

Consider a time series of dyadic length, $z_t, t = 1, \ldots, n$, where $n = 2^J$. The discrete wavelet transformation (DWT) decomposes the time series into J wavelet coefficients vectors, $W_j, j = 0, \ldots, J-1$ each of length $n_j = 2^{J-j}, j = 1, \ldots, J$ plus a

scaling coefficient V_J. Each wavelet coefficient is constructed as a difference of two weighted averages each of length $\lambda_j = 2^{j-1}$. Like the discrete Fourier transformation, the DWT provides an orthonormal decomposition, $W = \mathcal{W}Z$, where $W' = (W_1', \ldots, W_{J-1}', V_{J-1}')$, $Z = (z_1, \ldots, z_n)'$, and $\mathcal{W}$ is an orthonormal matrix. In practice, the DWT is not computed using matrix multiplication but much more efficiently using filtering and downsampling (Percival and Walden, 2000, Chapter 4). The resulting algorithm is known as the pyramid algorithm, and computationally, it is even more efficient than the fast Fourier transform. Applying the operations in reverse order yields the inverse DWT. Sometimes, a partial transformation is done, producing the wavelet coefficient vectors W_j, $j = 0, \ldots, J_0$, where $J_0 < J - 1$. In this case, the scaling coefficients are in the vector, V_{J_0} of length 2^{J-J_0}. The wavelet coefficients are associated with changes in the time series over the scale $\lambda_j = 2^{j-1}$, while the scaling coefficients, V_{J_0}, are associated with the average level on scale $\tau = 2^{J_0}$. The maximum overlap DWT or MODWT omits the downsampling. The MODWT has many advantages over the DWT (Percival and Walden, 2000, Chapter 5), even though it does not provide an orthogonal decomposition. Percival and Walden (2000) provide an extensive treatment of wavelet methods for time series research with many interesting scientific time series. Gençay et al. (2002) follows a similar approach to wavelets as given by Percival and Walden (2000) but with an emphasis on financial and economic applications.

All important methods as well as all datasets discussed in the books by Percival and Walden (2000) and Gençay et al. (2002) are available in the R packages `waveslim` (Whitcher, 2010) and `wmtsa` (Constantine and Percival, 2010). Nason (2008) provides a general introduction to wavelet methods in statistics, including smoothing and multiscale time series analysis. R scripts are used extensively in his book and all figures in the book (Nason, 2008) may be reproduced using R scripts available in the `wavethresh` R package (Nason, 2010).

Figure 25 shows the denoised annual Nile riverflows (Hipel and McLeod, 1994) using the universal threshold with hard thresholding and Haar wavelets. (Hipel and McLeod, 1994; Hipel et al., 1975) fit a step intervention analysis time series model with AR(1) noise. Physical reasons as well as cumsum analysis were presented (Hipel and McLeod, 1994, Section 19.2.4) to suggest 1903 as the start of intervention that was due to the operation of the Aswan dam. The fitted step intervention is represented by the three line segments, whereas the denoised flows are represented by the jagged curve. The points show actual observed flows. Figure 25 suggests the intervention actually may have started a few years prior to 1903. The computations for Fig. 25 were done using the functions `modwt()`, `universal.thresh.modwt()` and `imodwt()` in the package `waveslim`.

An estimate of the wavelet variance, $\hat{\sigma}^2(\lambda_j)$, is obtained based on the variance of the wavelet coefficients in an MODWT transformation at scale $\lambda_j = 2^{j-1}$. The wavelet variance is closely related to the power spectral density function and

$$\hat{\sigma}^2(\lambda_j) \approx 2 \int_{1/\lambda_j}^{2/\lambda_j} p(f) df.$$

The wavelet variance decomposition for the annual sunspot numbers, `sunspot.year` in R is shown in Fig. 26. This figure was produced using the `wavVar` function in `wmtsa`

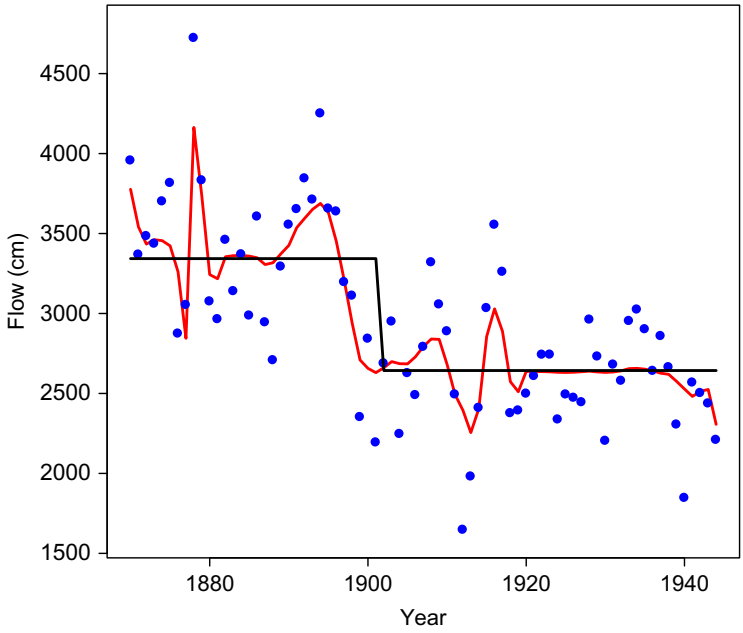

Fig. 25. Mean annual Nile flow, October to September, Aswan.

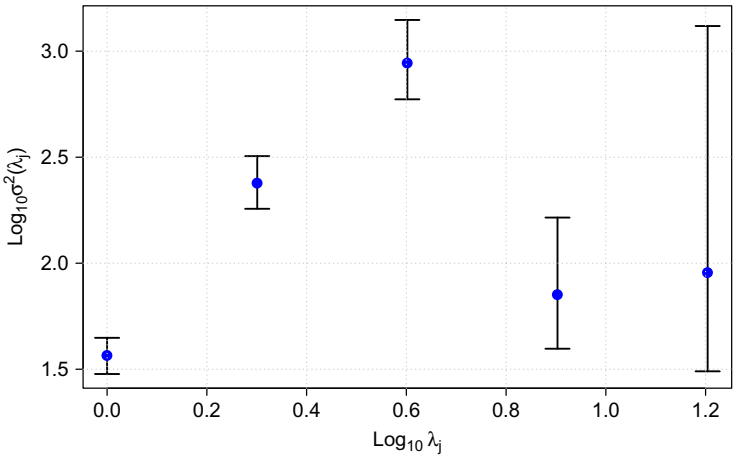

Fig. 26. Wavelet variance, yearly sunspot numbers, 1700–1988.

and the associated plot method. The 95% confidence intervals are shown in Fig. 26. The wavelet variances correspond to changes over 1, 2, 4, 8, and 16 years.

Multiresolution analysis (MRA) is another widely useful wavelet method for time series analysis. The MRA decomposition works best with the MODWT. The `mra` function in `waveslim` was used to produce the decomposition of an electrocardiogram time series that is shown in Fig. 27. The `la8` or least-asymmetric filter with half-length 8

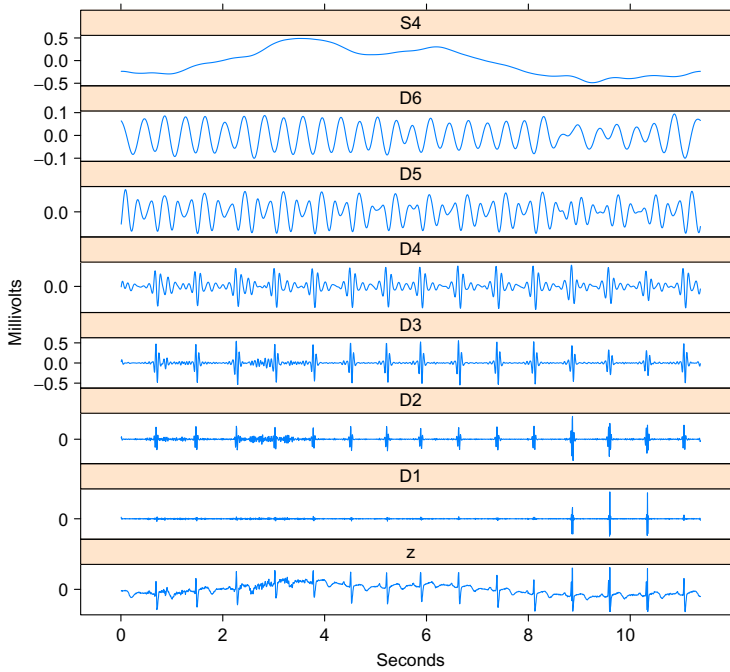

Fig. 27. MRA using MODWT with `la8` filter. ECG time series comprised of about 15 beats of a human heart, sampled at 180 Hz, units are millivolts and $n = 2048$.

was used (Percival and Walden, 2000, p. 109). A similar plot is given by Percival and Walden (2000, Fig. 184).

10. Stochastic differential equations (SDEs)

A SDE is comprised of a differential equation that includes a stochastic process, the simplest example being Brownian motion. Geometrical Brownian motion is often used to describe stock market prices. This SDE may be written, $dP(t) = P(t)\mu\, dt + P(t)\sigma\, dW(t)$, where $P(t)$ is the price at time t and the parameters $\mu > 0$ and $\sigma > 0$ are the drift and diffusion parameters. The Gaussian white noise term, $W(t)$, may be considered the derivative of Brownian motion. This SDE may also be written, $d\log(P(t)) = \mu\, dt + \sigma\, dW(t)$, so we see that $P(t) > 0$ and $\log(P(t))$ is Brownian motion.

More complicated SDE's may involve more complex drift and volatility functions. The book (Iacus, 2008) provides an intuitive and informal introduction to SDE and could be used in an introductory course on SDE. Only SDE's with Gaussian white noise are considered. The accompanying R package (Iacus, 2009) provides R scripts for all figures in the book (Iacus, 2008) as well as functions for simulation and statistical inference with SDE.

An important area of application is in financial mathematics, where option values or risk assessments are often driven by SDE systems. Usually, Monte-Carlo simulation is

the only way to find approximate solutions. The main class of SDE considered by this package is a diffusion process of the following form,

$$dX(t) = b(t, X(t))dt + \sigma(t, X(t))dW(t) \tag{16}$$

with some initial condition $X(0)$, where $W(t)$ is a standard Brownian motion. According to Itô formula, (16) can be represented as

$$X(t) = X(0) + \int_0^t b(u, X(u))du + \int_0^t \sigma(u, X(u))dW(u).$$

Under some regular conditions on the drift $b(\cdot,\cdot)$ and diffusion $\sigma^2(\cdot,\cdot)$, (16) has either a unique strong or weak solution. In practice, the class of SDE given by (16) is too large. The following diffusion process covers many well-known and widely used stochastic processes, including Vasicek (VAS), Ornstein-Uhlenbeck (OU), Black-Scholes-Merton (BS) or geometric Brownian motion, and Cox-Ingersoll-Ross (CIR),

$$dP(t) = P(t)\mu\, dt + P(t)\sigma dW(t) dX(t) = b(X(t))dt + \sigma(X(t))dW(t). \tag{17}$$

The main function is `sde.sim()`, and it has extensive options for the general diffusion process (17) or more specific processes. The function `DBridge()` provides another general purpose function for simulating diffusion bridges. Simple to use functions for simulating a Brownian bridge and geometric Brownian motion, `BBridge()`, and `GBM()` are also provided. Using `sde.sim()`, we simulate ten replications of Brownian motions each starting at the $X(0) = 0$ and comprised of 1000 steps. The results are displayed in Fig. 28.

A more complex SDE,

$$dX(t) = (5 - 11x + 6x^2 - x^3)dt + dW(t)$$

with $X(0) = 5$ is simulated using three different algorithms and using two different step sizes $\Delta = 0.1$ and $\Delta = 0.25$. For the smaller step size $\Delta = 0.1$, Fig. 29 suggests all three algorithms work about equally well. But only the Shoji-Ozaki algorithm appears to work with the larger step size $\Delta = 0.25$.

In addition to simulation, the `sde` package provides functions for parametric and nonparametric estimation: `EULERloglik()`, `ksmooth()`, `SIMloglik()`, and `simple.ef()`. Approximation of conditional density $X(t)|X(t_0) = x_0$ at point x_0 of a diffusion process is available with the functions: `dcElerian()`, `dcEuler()`, `dcKessler()`, `dcozaki()`, `dcShoji()`, and `dcSim()`.

11. Conclusion

There are many more packages available for time series than discussed in this article and many of these are briefly described in the CRAN Task Views.[10] In particular, see task views for **Econometrics**, **Finance**, and **TimeSeries**. We have selected those

[10] http://cran.r-project.org/web/views/

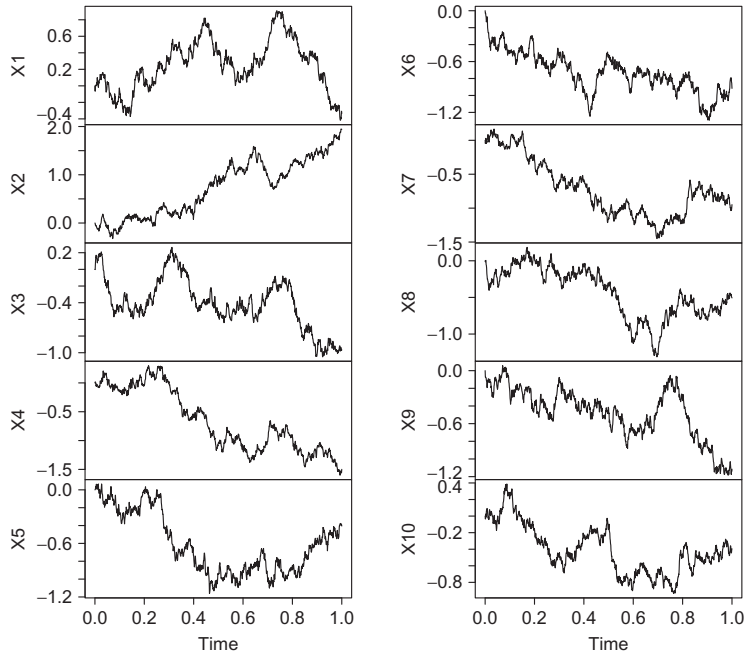

Fig. 28. Ten Brownian motions.

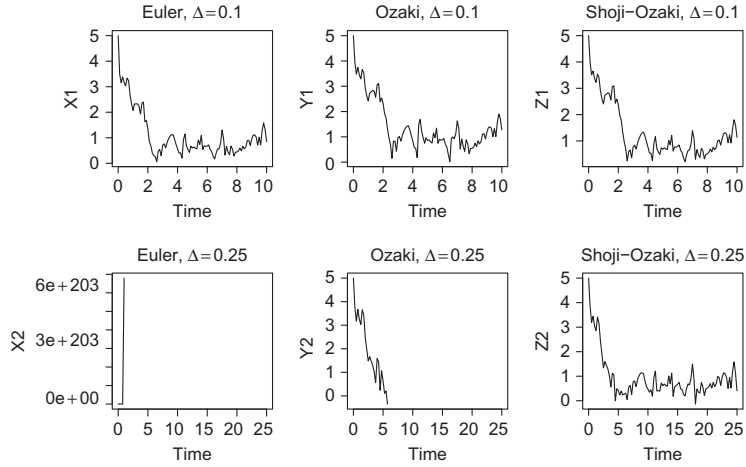

Fig. 29. Simulations of $dX(t) = (5 - 11x + 6x^2 - x^3)dt + dW(t)$ using three different algorithms and two different step sizes.

packages that might be of most general interest that have been most widely used and that we are most familiar with. The reader should note that the packages published on CRAN, including those in the task views, need only obey formatting rules and not produce computer errors. There is no endorsement that packages available on

CRAN produce correct or useful results. On the other hand, packages discussed in the *Journal of Statistical Software* or published by major publishers such as Springer-Verlag or Chapman & Hall/CRC have been carefully reviewed for correctness and quality.

Researchers wishing to increase the impact of their work should consider implementing their methods in R and making it available as a package on CRAN. Developing R packages is discussed in the online publication by R Development Core Team (2011) and from a broader perspective by Chambers (2008).

Acknowledgments

Drs. A. I. McLeod and Hao Yu would like to thank NSERC for Discovery Grants awarded to each of us. The authors would also like to thank Achim Zeileis for some suggestions and an anonymous referee for their comments.

A. Appendix

A.1. Datasets

Table A.1
Datasets in the packge 'datasets'

Dataset name	Description
`AirPassengers`	Monthly airline passengers, 1949–1960
`BJsales`	Sales data with leading indicator
`BOD`	Biochemical oxygen demand
`EuStockMarkets`	Daily close price, European stocks, 1991–1998
`LakeHuron`	Level of Lake Huron 1875–1972
`Nile`	Flow of the river Nile
`UKDriverDeaths`	Road casualties, Great Britain 1969–1984
`UKgas`	UK quarterly gas consumption
`USAccDeaths`	Accidental deaths in the US 1973–1978
`USPersonalExpenditure`	Personal expenditure data
`WWWusage`	Internet usage per minute
`WorldPhones`	Ihe world's telephones
`airmiles`	Passenger miles, US airlines, 1937–1960
`austres`	Quarterly time series, Australian residents
`co2`	Mauna loa atmospheric co2 concentration
`UKLungDeaths`	Monthly deaths from lung diseases in the UK
`freeny`	Freeny's revenue data
`longley`	Longley's economic regression data
`lynx`	Annual Canadian lynx trappings 1821–1934
`nhtemp`	Average yearly temperatures in New Haven
`nottem`	Monthly temperature, Nottingham, 1920–1939
`sunspot.month`	Monthly sunspot data, 1749–1997
`sunspot.year`	Yearly sunspot data, 1700–1988
`sunspots`	Monthly sunspot numbers, 1749–1983
`treering`	Yearly treering data, -6000–1979
`uspop`	Populations recorded by the US census

A.2. **stats**

Table A.2
'stats' package utilities for ts objects. These functions are useful for creating and manipulating univariate and multivariate time series

Function	Purpose
`embed`	Matrix containing lagged values
`lag`	Lagged values
`ts`	Create a time series object
`ts.intersect`	Intersection, multivariate series by
`ts.union`	Union, multivariate series by union
`time`	Extract time from a ts object
`cycle`	Extract seasonal times from a ts object
`frequency`	Sampling interval
`window`	Select subset of time series

Table A.3
'stats' package autocorrelation and spectral analysis functions

Function	Purpose
`acf`	acf, pacf
`ccf`	Cross-correlation
`cpgram`	Bartlett's cumulate periodogram test
`lag.plot`	Alternative time series plot
`fft`	Fast Fourier transform
`convolve`	Convolution via fft
`filter`	Moving-average/autoregressive filtering
`spectrum`	Spectral density estimation
`toeplitz`	Toeplitz matrix

Table A.4
'stats' package functions for time series models. In addition, many of these functions have `predict` and `residuals` methods

Function	Purpose
`arima, arima0`	Fit ARIMA
`ar`	Fit AR
`KalmanLike`	Log-likelihood, univariate state-space model
`KalmanRun`	KF filtering
`KalmanSmooth`	KF smoothing
`KalmanForecast`	KF forecasting
`makeARIMA`	ARIMA to KF
`PP.test`	Phillips-Perron unit-root test
`tsdiag`	Diagnostic checks
`ARMAacf`	Theoretical ACF of ARMA
`acf2AR`	Fit AR to ACF
`Box.test`	Box-Pierce or Ljung-Box test
`diff, diffinv`	Difference or inverse
`ARMAtoMA`	MA expansion for ARMA
`arima.sim`	Simulate ARIMA
`HoltWinters`	Holt-Winters filtering
`StructTS`	Kalman filter modeling

Table A.5
'stats' package smoothing and filtering

Function	Purpose
`filter`	Moving-average/autoregressive filtering
`tsSmooth`	Smooth from StuctTS object
`stl`	Seasonal-trend-Loess decomposition
`decompose`	Seasonal decomposition, moving-average filters

A.3. **tseries**

Table A.6
'tseries' package functions

Function	Purpose
`adf.test`	Augmented Dickey-Fuller test
`bds.test`	Breusch-Godfrey test
`garch`	Fit GARCH models to time series
`get.hist.quote`	Download historical finance data
`jarque.bera.test`	Jarque-Bera test
`kpss.test KPSS`	Test for stationarity
`quadmap`	Quadratic map (logistic equation)
`runs.test`	Runs test
`terasvirta.test`	Teraesvirta neural network test for nonlinearity
`tsbootstrap`	Bootstrap for general stationary data
`white.test`	White neural network test for nonlinearity

Table A.7
'tseries' package datasets

Dataset name	Description
`bev`	Beveridge wheat price index, 1500-1869
`camp`	Mount Campito, treering data, -3435-1969
`ice.river`	Icelandic river Data
`NelPlo`	Nelson-Plosser macroeconomic time series
`nino`	sea surface temperature, El Niño indices
`tcm`	monthly yields on treasury securities
`tcmd`	daily yields on treasury securities
`USeconomic`	US economic variables

A.4. **'Forecast'** Package

Table A.8
General purpose utility functions

Function	Purpose
`accuracy()`	Accuracy measures of forecast
`BoxCox, invBoxCox()`	Box-Cox transformation
`decompose()`	Improved version of decompose()

(Continued)

Table A.8
General purpose utility functions (*Continued*)

Function	Purpose
`dm.test()`	Diebold-Mariano test compares the forecast accuracy
`forecast()`	Generic function with various methods
`monthdays()`	Number of days in seasonal series
`na.interp()`	Interpolate missing values
`naive(), snaive()`	ARIMA(0,1,0) forecast and seasonal version
`seasadj()`	Seasonally adjusted series
`seasonaldummy()`	Create matrix of seasonal indicator variables
`seasonplot()`	Season plot

Table A.9
ARIMA functions

Function	Purpose
`arfima`	Automatic ARFIMA
`Arima`	Improved version of arima()
`arima.errors`	Removes regression component
`auto.arima`	Automatic ARIMA modeling
`ndiffs`	Use unit-root test to determine differencing
`tsdisplay()`	Display with time series plot, ACF, PACF, etc.

Table A.10
Exponential smoothing and other time series modeling functions

Function	Purpose
`croston`	Exponential forecasting for intermittent series
`ets`	Exponential smoothing state-space model
`logLik.ets`	Loglikelihood for ets object
`naive(), snaive()`	ARIMA(0,1,0) forecast and seasonal version
`rwf()`	Random walk forecast with possible drifts
`ses(), holt(), hw()`	Exponential forecasting methods
`simulate.ets()`	Simulation method for ets object
`sindexf`	Seasonal index, future periods
`splinef`	Forecast using splines
`thetaf`	Forecast using theta method
`tslm()`	lm()-like function using trend and seasonal

A.5. **ltsa**

Table A.11
Main functions in **ltsa**

Function	Purpose
`DHSimulate`	Simulate using Davies-Harte method
`DLLoglikelihood`	Exact concentrated log-likelihood
`DLResiduals`	Standardized prediction residuals

(Continued)

Table A.11
Main functions in **ltsa** (*Continued*)

Function	Purpose
`DLSimulate`	Simulate using DL recursion
`SimGLP`	Simulate general linear process
`TrenchInverse`	Toeplitz matrix inverse
`ToeplitzInverseUpdate`	Updates the inverse
`TrenchMean`	Exact MLE for mean
`TrenchForecast`	Exact forecast and variance

A.6. **FitAR**

Table A.12
FitAR model selection functions

Function	Purpose
`PacfPlot`	Partial autocorrelation plot
`SelectModel`	AIC/BIC selection
`TimeSeriesPlot`	Time series plot

Table A.13
FitAR estimation functions

Function	Purpose
`FitAR`	Exact mle for AR(p)/subset ARzeta
`FitARLS`	LS for AR(p)/subset ARphi
`GetFitAR`	Fast exact mle for AR(p)/subset ARzeta
`GetFitARLS`	Fast LS for AR(p) and subset ARphi
`GetARMeanMLE`	Exact mean MLE in AR
`AR1Est`	Exact MLE for mean-zero AR(1)

Table A.14
FitAR diagnostic check functions

Function	Purpose
`Boot`	Generic parametric bootstrap
`Boot.FitAR`	Method for FitAR
`Boot.ts`	Method for ts
`LjungBox`	Ljung-Box portmanteau test
`LBQPlot`	Plot Ljung-Box test results
`RacfPlot`	Residual acf plot
`JarqueBeraTest`	Test for normality

Table A.15
FitAR miscellaneous functions

Function	Purpose
`AcfPlot`	General purpose correlation plotting
`ARSdf`	AR spectral density via FFT
`ARToMA`	Impulse coefficients
`ARToPacf`	Transform AR to PACF
`BackcastResidualsAR`	Compute residuals using backforecasting
`cts`	Concantenate time series
`InformationMatrixAR`	Fisher information matrix AR
`InformationMatrixARp`	Fisher information matrix subset case, ARp
`InformationMatrixARz`	Fisher information matrix subset case, ARz
`InvertibleQ`	Test if invertible or stationary-casual
`PacfDL`	Compute PACF from ACF using DL recursions
`PacfToAR`	Transform PACF to AR
`sdfplot`	Generic spectral density plot
`sdfplot.FitAR`	Method for class FitAR
`sdfplot.Arima`	Method for class Arima
`sdfplot.ar`	Method for class ar
`sdfplot.ts`	Method for class ts
`sdfplot.numeric`	Method for class numeric
`SimulateGaussianAR`	Simulate Gaussian AR
`Readts`	Input time series
`TacvfAR`	Theoretical autocovariances AR
`TacvfMA`	Theoretical autocovariances MA
`VarianceRacfAR`	Variance of residual acf, AR
`VarianceRacfARp`	Variance of residual acf, subset case, ARp
`VarianceRacfARz`	Variance of residual acf, subset case, ARz

References

Adler, J., 2009. R in a Nutshell. O'Reilly, Sebastopol, CA.

Aknouche, A., Bibi, A., 2009. Quasi-maximum likelihood estimation of periodic garch and periodic arma-garch processes. J. Time Ser. Anal. 30(1), 19–46.

Baillie, R.T., 1996. Long memory processes and fractional integration in econometrics. J. Econom. 73(1), 5–59.

Banerjee, A., Dolado, J.J., Galbraith, J.W., Hendry, D.F., 1993. Cointegration, Error Correction, and the Econometric Analysis of Non-Stationary Data. Oxford University Press, Oxford.

Becker, R.A., Clark, L.A., Lambert, D., 1994. Cave plots: a graphical technique for comparing time series. J. Comput. Graph. Stat. 3(3), 277–283.

Beran, J., 1994. Statistics for Long Memory Processes. Chapman & Hall/CRC, Boca Raton.

Beran, J., Whitcher, B., Maechler, M., 2009. Longmemo: Statistics for Long-Memory Processes. http://CRAN.R-project.org/package=longmemo (accessed 03.02.12).

Bloomfield, P., 2000. Fourier Analysis of Time Series: An Introduction, second ed. Wiley, New York.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econom. 31(3), 307–327.

Box, G., Jenkins, G.M., Reinsel, G.C., 2008. Time Series Analysis: Forecasting and Control, fourth ed. Hoboken, N.J., Wiley, New York.

Box, G.E.P., Tiao, G.C., 1975. Intervention analysis with applications to economic and environmental problems. J. Am. Stat. Assoc. 70(349), 70–79.

Braun, W.J., Murdoch, D.J., 2008. A First Course in Statistical Programming with R. Cambridge University Press, Cambridge.
Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods, second ed. Springer, New York.
Brown, R.L., Durbin, J., Evans, J.M., 1975. Techniques for testing the constancy of regression relationships over time. J. R. Stat. Soc. B 37, 149–163.
Chambers, J.M., June 2008. Software for Data Analysis: Programming with R. Statistics and Computing. Springer-Verlag, New York.
Chan, K.-S., 2011. TSA: Time Series Analysis. R package version 0.98. http://CRAN.R-project.org/package=TSA (accessed 03.02.12).
Chan, N.H., 2010. Time Series: Applications to Finance with R, third ed. Wiley, New York.
Cleveland, W.S., 1993. Visualizing Data. Hobart Press, Summit, New Jersey.
Cleveland, W.S., McGill, M.E., McGill, R., 1988. The shape parameter of a two-variable graph. J. Am. Stat. Assoc. 83(402), 289–300.
Constantine, W., Percival, D., 2010. wmtsa: Insightful Wavelet Methods for Time Series Analysis. R package version 1.0-5. http://CRAN.R-project.org/package=wmtsa (accessed 03.02.12).
Cook, D., Swayne, D.F., 2007. Interactive Dynamic Graphics for Data Analysis R. Springer-Verlag, New York.
Cowpertwait, P.S., Metcalfe, A.V., 2009. Introductory Time Series with R. Springer Science+Business Media, LLC, New York.
Craigmile, P.F., 2003. Simulating a class of stationary gaussian processes using the davies-harte algorithm, with application to long memory processes. J. Time Ser. Anal. 24, 505–511.
Crawley, M.J., 2007. The R Book. Wiley, New York.
Cryer, J.D., Chan, K.-S., 2008. Time Series Analysis: With Applications in R, second ed. Springer Science+Business Media, LLC, New York.
Dalgaard, P., 2008. Introductory Statistics with R. Springer Science+Business Media, LLC, New York.
Davies, R.B., Harte, D.S., 1987. Tests for hurst effect. Biometrika 74, 95–101.
Dethlefsen, C., Lundbye-Christensen, S., Christensen, A.L., 2009. sspir: State Space Models in R. http://CRAN.R-project.org/package=sspir (accessed 03.02.12).
Diebold, F.X., Rudebusch, G.D., 1989. Long memory and persistence in aggregate output. J. Monet. Econ. 24, 189–209.
Durbin, J., Koopman, S.J., 2001. Time Series Analysis by State Space Methods. Oxford University Press, Oxford.
Elliott, G., Rothenberg, T.J., Stock, J.H., 1996. Efficient tests for an autoregressive unit root. Econometrica 64(4), 813–836.
Enders, W., 2010. Applied Econometric Time Series, third ed. John Wiley and Sons, New York.
Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica 50(4), 987–1007.
Engle, R.F., Granger, C.W.J., 1987. Co-integration and error correction: representation, estimation, and testing. Econometrica 55(2), 251–276.
Everitt, B.S., Hothorn, T., 2009. A Handbook of Statistical Analyses Using R, second ed. Chapman and Hall/CRC, Boca Raton.
Faraway, J., Chatfield, C., 1998. Time series forecasting with neural networks: a comparative study using the airline data. J. R. Stat. Soc. Ser. C (Appl. Stat.) 47(2) 231–250.
Fox, J., 2005. The R commander: A basic-statistics graphical user interface to R. J. Stat Software 14(9), 1–42. http://www.jstatsoft.org/v14/i09 (accessed 03.02.12).
Fraley, C., Leisch, F., Maechler, M., Reisen, V., Lemonte, A., 2009. fracdiff: Fractionally differenced ARIMA aka ARFIMA(p,d,q) models. http://CRAN.R-project.org/package=fracdiff (accessed 03.02.12).
Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19(1), 1–67.
Gelper, S., Fried, R., Croux, C., 2010. Robust forecasting with exponential and holt-winters smoothing. J. Forecast. 29(3), 285–300.
Gençay, R., Selçuk, F., Whitcher, B., 2002. An Introduction to Wavelets and Other Filtering Methods in Finance and Economics. Academic Press, New York.
Gentleman, R., 2009. R Programming for Bioinformatics. Chapman and Hall/CRC, Boca Raton.
Gentleman, R., Ihaka, R., 1996. R: A language for data analysis and graphics. J. Comput. Graph. Stat. 5(2), 491–508.

Gilbert, P., 1993. State space and arma models: An overview of the equivalence. Bank of Canada Publications. Working Paper 1993-00. http://www.bankofcanada.ca/1993/03/publications/research/working-paper-199/ (accessed 03.02.12).

Gilbert, P., 2011. dse: Dynamic Systems Estimation (time series package). http://CRAN.R-project.org/package=dse (accessed 03.02.12).

Granger, C.W.J., Newbold, P., 1974. Spurious regressions in econometrics. J. Econom. 2, 111–120.

Granger, C.W.J., Newbold, P., 1976. Forecasting transformed series. J. R. Stat. Soc. Ser. B (Methodological) 38(2), 189–203.

Graves, S., 2011. FinTS: Companion to Tsay (2005) Analysis of Financial Time Series. R package version 0.4-4. http://CRAN.R-project.org/package=FinTS (accessed 03.02.12).

Grolemund, G., Wickham, H., 2011. Dates and times made easy with lubridate. J. Stat. Soft. 40(3), 1–25. http://www.jstatsoft.org/v40/i03 (accessed 03.02.12).

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton, NJ.

Hansen, B.E., 1995. Rethinking the univariate approach to unit root testing: using covariates to increase power. Econom. Theory 11(5), 1148–1171.

Harrison, J., West, M., 1997. Bayesian Forecasting and Dynamic Models. Springer, New York.

Harvey, A., 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge.

Haslett, J., Raftery, A.E., 1989. Space-time modelling with long-memory dependence: assessing Ireland's wind power resource. J. R. Stat. Soc. Ser. C (Appl. Stat.) 38(1), 1–50.

Hastie, T., Tibshirani, R., 2011. mda: Mixture and Flexible Discriminant Analysis. R package version 0.4-2. http://CRAN.R-project.org/package=mda (accessed 03.02.12).

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning, second ed. Springer-Verlag, New York.

Heiberger, R.M., Neuwirth, E., 2009. R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics. Springer Science+Business Media, LLC, New York.

Helske, J., 2011. KFAS: Kalman filter and smoothers for exponential family state space models. http://CRAN.R-project.org/package=KFAS (accessed 03.02.12).

Hipel, K.W., Lennox, W.C., Unny, T.E., McLeod, A.I., 1975. Intervention analysis in water resources. Water Resour. Res. 11(6), 855–861.

Hipel, K.W., McLeod, A.I., 1994. Time Series Modelling of Water Resources and Environmental Systems. Elsevier, Amsterdam.

Hoffmann, T.J., 2011. Passing in command line arguments and parallel cluster/multicore batching in r with batch. J. Stat. Soft., Code Snippets 39(1), 1–11. http://www.jstatsoft.org/ -v39/c01 (accessed 03.02.12).

Hopwood, W.S., McKeown, J.C., Newbold, P., 1984. Time series forecasting models involving power transformations. J. Forecast. 3(1), 57–61.

Hornik, K., Leisch, F., 2001. Neural network models. In: Peña, D., Tiao, G.C., Tsay, R.S. (Eds.), A Course in Time Series Analysis. Wiley, New York, Ch. 13, pp. 348–364.

Hothorn, T., Zeileis, A., Millo, G., Mitchell, D., 2010. lmtest: Testing Linear Regression Models. R package version 0.9-27. http://CRAN.R-project.org/package=lmtest (accessed 03.02.12).

Hyndman, R.J., 2010. forecast: Forecasting functions for time series. R package version 2.17. http://CRAN.R-project.org/package=forecast (accessed 03.02.12).

Hyndman, R.J., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. J. Stat. Softw. 27(3), 1–22. http://www.jstatsoft.org/v27/i03 (accessed 03.02.12).

Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D., 2008. Forecasting with Exponential Smoothing: The State Space Approach. Springer-Verlag, New York.

Iacus, S.M., 2008. Simulation and Inference for Stochastic Differential Equations: With R Examples. Springer Science+Business Media, LLC, New York.

Iacus, S.M., 2009. sde: Simulation and Inference for Stochastic Differential Equations. R package version 2.0.10. http://CRAN.R-project.org/package=sde (accessed 03.02.12).

Johansen, S., 1995. Likelihood-Based Inference in Cointegrated Vector Autoregressive Models. Oxford University Press, Oxford.

Kajitani, Y., Hipel, K.W., McLeod, A.I., 2005. Forecasting nonlinear time series with feed-forward neural networks: a case study of canadian lynx data. J. Forecast. 24, 105–117.

Kedem, B., Fokianos, K., 2002. Regression Models for Time Series Analysis. Wiley, New York.

Keenan, D.M., 1985. A Tukey nonadditivity-type test for time series nonlinearity. Biometrika 72, 39–44.
Kleiber, C., Zeileis, A., 2008. Applied Econometrics with R. Springer, New York.
Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? J. Econom. 54, 159–178.
Lee, T.H., White, H., Granger, C.W.J., 1993. Testing for neglected nonlinearity in time series models. J. Econom. 56, 269–290.
Leisch, F., 2002. Dynamic generation of statistical reports using literate data analysis. In: Härdle, W., Rönz, B. (Eds.), COMPSTAT 2002 – Proceedings in Computational Statistics. Physica-Verlag, Heidelberg, pp. 575–580.
Leisch, F., 2003. Sweave and beyond: computations on text documents. In: Hornik, K., Leisch, F., Zeileis, A. (Eds.), Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria. ISSN 1609-395X. http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/ (accessed 03.02.12).
Lewis, P.A.W., Stevens, J.G., 1991. Nonlinear modeling of time series using multivariate adaptive regression splines (mars). J. Am. Stat. Assoc. 86(416), 864–877.
Li, W.K., 1994. Time series models based on generalized linear models: some further results. Biometrics 50(2), 506–511.
Luethi, D., Erb, P., Otziger, S., 2010. FKF: Fast Kalman Filter. http://CRAN.R-project.org/package=FKF (accessed 03.02.12).
Lupi, C., 2011. CADFtest: Hansen's Covariate-Augmented Dickey-Fuller Test. R package version 0.3-1. http://CRAN.R-project.org/package=CADFtest (accessed 03.02.12).
Lütkepohl, H., 2005. New Introduction to Multiple Time Series Analysis. Springer-Verlag, New York.
Lütkepohl, H., Krätzig, M. (Eds.), 2004. Applied Time Series Econometrics. Cambridge University Press, Cambridge.
MacKinnon, J.G., 1996. Numerical distribution functions for unit root and cointegration tests. J. Appl. Econom. 11, 601–618.
McLeod, A.I., 1994. Diagnostic checking periodic autoregression models with application. J. Time Ser. Anal. 15, 221–223, Addendum, J. Time Ser. Anal. 16, 647–648.
McLeod, A.I., 1998. Hyperbolic decay time series. J. Time Ser. Anal. 19, 473–484.
McLeod, A.I., 2010. FitARMA: Fit ARMA or ARIMA Using Fast MLE Algorithm. R package version 1.4. http://CRAN.R-project.org/package=FitARMA (accessed 03.02.12).
McLeod, A.I., Balcilar, M., 2011. pear: Package for Periodic Autoregression Analysis. R package version 1.2. http://CRAN.R-project.org/package=pear (accessed 03.02.12).
McLeod, A.I., Li, W.K., 1983. Diagnostic checking arma time series models using squared-residual autocorrelations. J. Time Ser. Anal. 4, 269–273.
McLeod, A.I., Yu, H., Krougly, Z., 2007. Algorithms for linear time series analysis: With R package. J. Stat. Softw. 23(5), 1–26. http://www.jstatsoft.org/v23/i05 (accessed 03.02.12).
McLeod, A.I., Yu, H., Krougly, Z., 2011a. FGN: Fractional Gaussian Noise, estimation and simulation. R package version 1.4. http://CRAN.R-project.org/package=ltsa (accessed 03.02.12).
McLeod, A.I., Zhang, Y., 2006. Partial autocorrelation parameterization for subset autoregression. J. Time Ser. Anal. 27(4), 599–612.
McLeod, A.I., Zhang, Y., 2008a. Faster arma maximum likelihood estimation. Comput. Stat. Data Anal. 52(4), 2166–2176.
McLeod, A.I., Zhang, Y., 2008b. Improved subset autoregression: With R package. J. Stat. Softw. 28(2), 1–28. http://www.jstatsoft.org/v28/i02 (accessed 03.02.12).
McLeod, A.I., Zhang, Y., Xu, C., 2011b. FitAR: Subset AR Model Fitting. R package version 1.92. http://CRAN.R-project.org/package=FitAR (accessed 03.02.12).
Meyer, D., June 2002. Naive time series forecasting methods: the holt-winters method in package ts. R News 2(2), 7–10.
Milborrow, S., 2011. earth: Multivariate Adaptive Regression Spline Models. R package version 2.6-2. http://CRAN.R-project.org/package=earth (accessed 03.02.12).
Moore, D.S., 2007. The Basic Practice of Statistics, fourth ed. W. H. Freeman & Co., New York.
Murrell, P., 2011. R Graphics, second ed. Chapman and Hall/CRC, Boca Raton.
Nason, G., 2008. Wavelet Methods in Statistics with R. Springer-Verlag, New York.
Nason, G., 2010. wavethresh: Wavelets statistics and transforms. R package version 4.5. http://CRAN.R-project.org/package=wavethresh (accessed 03.02.12).

Peng, R., 2008. A method for visualizing multivariate time series data. J. Stat. Softw. 25 (Code Snippet 1), 1–17. http://www.jstatsoft.org/v25/c01 (accessed 03.02.12).
Percival, D.B., Walden, A.T., 1993. Spectral Analysis For Physical Applications. Cambridge University Press, Cambridge.
Percival, D.B., Walden, A.T., 2000. Wavelet Methods for Time Series Analysis. Cambridge University Press, Cambridge.
Petris, G., 2010. dlm: Bayesian and Likelihood Analysis of Dynamic Linear Models. http://CRAN.R-project.org/package=dlm (accessed 03.02.12).
Petris, G., Petrone, S., Campagnoli, P., 2009. Dynamic Linear Models with R. Springer Science+Business Media, LLC, New York.
Pfaff, B., 2006. Analysis of Integrated and Cointegrated Time Series with R. Springer, New York.
Pfaff, B., 2008. Var, svar and svec models: implementation within R package vars. J. Stat. Softw. 27(4), 1–32. http://www.jstatsoft.org/v27/i04 (accessed 03.02.12).
Pfaff, B., 2010a. urca: Unit Root and Cointegration Tests for Time Series Data. R package version 1.2-5. http://CRAN.R-project.org/package=urca (accessed 03.02.12).
Pfaff, B., 2010b. vars: VAR Modelling. R package version 1.4-8. http://CRAN.R-project.org/package=vars (accessed 03.02.12).
Phillips, P.C.B., Ouliaris, S., 1990. Asymptotic properties of residual based tests for cointegration. Econometrica 58, 165–193.
Phillips, P.C.B., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika 75(2), 335–346.
R Development Core Team, 2011. Writing R Extensions. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/ (accessed 03.02.12).
Revolution Computing, 2011. foreach: For Each Looping Construct for R. R package version 1.3.2. http://CRAN.R-project.org/package=foreach (accessed 03.02.12).
Ripley, B.D., 2011. nnet: Feed-forward Neural Networks and Multinomial Log-Linear Models. R package version 7.3-1. http://CRAN.R-project.org/package=nnet (accessed 03.02.12).
Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, New York.
Ripley, B.D., June 2002. Time series in R 1.5.0. R News 2(2), 2–7.
Ritz, C., Streibig, J.C., 2008. Nonlinear Regression with R. Springer Science+Business Media, LLC, New York.
Said, S.E., Dickey, D.A., 1984. Test for unit roots in autoregressive-moving average models of unknown order. Biometrika 71(3), 599–607.
Sarkar, D., 2008. Lattice: Multivariate Data Visualization with R. Springer, New York.
Schmidberger, M., Morgan, M., Eddelbuettel, D., Yu, H., Tierney, L., Mansmann, U., Aug 2009. State of the art in parallel computing with R. J. Stat. Softw. 31(1), 1–27. http://www.jstat soft.org/v31/i01 (accessed 03.02.12).
Shadish, W.R., Cook, T.D., Campbell, D.T., 2001. Experimental and Quasi-Experimental Designs for Generalized Causal Inference, second ed. Houghton Mifflin, Boston.
Shumway, R.H., Stoffer, D.S., 2011. Time Series Analysis and Its Applications With R Examples, third ed. Springer, New York.
Smith, B., Field, C., 2001. Symbolic cumulant calculations for frequency domain time series. Stat. Comput. 11, 75–82.
Spector, P., 2008. Data Manipulation with R. Springer-Verlag, Berlin.
Teraesvirta, T., Lin, C.F., Granger, C.W.J., 1993. Power of the neural network linearity test. J. Time Ser. Anal. 14, 209–220.
Tesfaye, Y.G., Anderson, P.L., Meerschaert, M.M., 2011. Asymptotic results for fourier-parma time series. J. Time Ser. Anal. 32(2), 157–174.
Thompson, M.E., McLeod, A.I., June 1976. The effects of economic variables upon the demand for cigarettes in Canada. Math. Sci. 1, 121–132.
Trapletti, A., 2011. tseries: Time Series Analysis and Computational Finance. R package version 0.10-25. http://CRAN.R-project.org/package=tseries (accessed 03.02.12).
Tsay, R.S., 2010. Analysis of Financial Time Series, third ed. Wiley, New York.
Tusell, F., 2011. Kalman filtering in R. J. Stat. Softw. 39(2). http://www.jstatsoft.org/v39/i02 (accessed 03.02.12).
Ursu, E., Duchesne, P., 2009. On modelling and diagnostic checking of vector periodic autoregressive time series models. J. Time Ser. Anal. 30(1), 70–96.
Venables, W.N., Ripley, B.D., 2000. S Programming. Springer, New York.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, fourth ed. Springer, New York.
Vingilis, E., McLeod, A.I., Seeley, J., Mann, R.E., Stoduto, G., Compton, C., et al., 2005. Road safety impact of extended drinking hours in ontario. Accid. Anal. Prev. 37, 547–556.
Whitcher, B., 2010. waveslim: Basic Wavelet Routines for One-, Two- and Three-Dimensional Signal Processing. R package version 1.6.4. http://CRAN.R-project.org/package=waveslim (accessed 03.02.12).
Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer, New York.
Wilkinson, L., 1999. The Grammar of Graphics. Springer, New York.
Wolfram Research, Inc., 2011. Mathematica Edition: Version 8.0. Wolfram Research, Inc., Champaign, Illinois.
Wood, S., 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC, Boca Raton.
Wuertz, D., 2010. fBasics: Rmetrics - Markets and Basic Statistics. R package version 2110.79. http://CRAN.R-project.org/package=fBasics (accessed 03.02.12).
Wuertz, D., Chalabi, Y., 2011. timeSeries: Rmetrics - Financial Time Series Objects. R package version 2130.92. http://CRAN.R-project.org/package=timeSeries (accessed 03.02.12).
Wuertz, D., 2009a. fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling. R package version 2110.80. http://CRAN.R-project.org/package=fGarch (accessed 03.02.12).
Wuertz, D., 2009b. fUnitRoots: Trends and Unit Roots. R package version 2100.76. http://CRAN.R-project.org/package=fUnitRoots (accessed 03.02.12).
Würtz, D., 2004. Rmetrics: An Environment for Teaching Financial Engineering and Computational Finance with R. Rmetrics, ITP, ETH Zürich, Swiss Federal Institute of Technology, ETH Zürich, Switzerland. http://www.rmetrics.org (accessed 03.02.12).
Zeileis, A., 2010. dynlm: Dynamic Linear Regression. R package version 0.3-0. http://CRAN.R-project.org/package=dynlm (accessed 03.02.12).
Zeileis, A., Leisch, F., Hansen, B., Hornik, K., Kleiber, C., 2010. strucchange: Testing, Monitoring and Dating Structural Changes. R package version 1.4-4. http://CRAN.R-project.org/package=strucchange (accessed 03.02.12).
Zeileis, A., Leisch, F., Hornik, K., Kleiber, C., 2002. Strucchange: An R package for testing for structural change in linear regression models. J. Stat. Softw. 7(2), 1–38. http://www.jstatsoft.org/v07/i02 (accessed 03.02.12).
Zhang, Y., McLeod, A.I., 2006. Computer algebra derivation of the bias of burg estimators. J. Time Ser. Anal. 27, 157–165.
Zhou, L., Braun, W.J., 2010. Fun with the r grid package. J. Stat. Educ. 18. http://www.amstat.org/publications/jse/v18n3/zhou.pdf (accessed 03.02.12).
Zivot, E., Wang, J., 2006. Modeling Financial Time Series with S-PLUS, second Edition. Springer Science+Business Media, Inc, New York.
Zucchini, W., MacDonald, I.L., 2009. Hidden Markov Models for Time Series: A Practical Introduction using R, second ed. Chapman &Hall/CRC, Boca Raton.
Zuur, A.F., Ieno, E.N., Meesters, E., 2009. A Beginner's Guide to R. Springer Science+Business Media, LLC, New York.

Index

Page numbers followed by "*f*" indicates figures and "*t*" indicates tables.

D

E

F

G

H

M

R

S

T

U

V

W

This page intentionally left blank

Handbook of Statistics
Contents of Previous Volumes

Volume 1. Analysis of Variance
Edited by P.R. Krishnaiah
1980 xviii + 1002 pp.

Volume 2. Classification, Pattern Recognition and Reduction of Dimensionality
Edited by P.R. Krishnaiah and L.N. Kanal
1982 xxii + 903 pp.

27. White Blood Cell Recognition by F.S. Gelsema and G.H. Landweerd
28. Pattern Recognition Techniques for Remote Sensing Applications by P.H. Swain
29. Optical Character Recognition – Theory and Practice by G. Nagy
30. Computer and Statistical Considerations for Oil Spill Identification by Y.T. Chien and T.J. Killeen
31. Pattern Recognition in Chemistry by B.R. Kowalski and S. Wold
32. Covariance Matrix Representation and Object-Predicate Symmetry by T. Kaminuma, S. Tomita and S. Watanabe
33. Multivariate Morphometrics by R.A. Reyment
34. Multivariate Analysis with Latent Variables by P.M. Bentler and D.G. Weeks
35. Use of Distance Measures, Information Measures and Error Bounds in Feature Evaluation by M. Ben-Bassat
36. Topics in Measurement Selection by J.M. Van Campenhout
37. Selection of Variables Under Univariate Regression Models by P.R. Krishnaiah
38. On the Selection of Variables Under Regression Models Using Krishnaiah's Finite Intersection Tests by J.L. Schmidhammer
39. Dimensionality and Sample Size Considerations in Pattern Recognition Practice by A.K. Jain and B. Chandrasekaran
40. Selecting Variables in Discriminant Analysis for Improving upon Classical Procedures by W. Schaafsma
41. Selection of Variables in Discriminant Analysis by P.R. Krishnaiah

Volume 3. Time Series in the Frequency Domain
Edited by D.R. Brillinger and P.R. Krishnaiah
1983 xiv + 485 pp.

1. Wiener Filtering (with emphasis on frequency-domain approaches) by R.J. Bhansali and D. Karavellas
2. The Finite Fourier Transform of a Stationary Process by D.R. Brillinger
3. Seasonal and Calendar Adjustment by W.S. Cleveland
4. Optimal Inference in the Frequency Domain by R.B. Davies
5. Applications of Spectral Analysis in Econometrics by C.W.J. Granger and R. Engle
6. Signal Estimation by E.J. Hannan
7. Complex Demodulation: Some Theory and Applications by T. Hasan
8. Estimating the Gain of a Linear Filter from Noisy Data by M.J. Hinich
9. A Spectral Analysis Primer by L.H. Koopmans
10. Robust-Resistant Spectral Analysis by R.D. Martin
11. Autoregressive Spectral Estimation by E. Parzen
12. Threshold Autoregression and Some Frequency-Domain Characteristics by J. Pemberton and H. Tong
13. The Frequency-Domain Approach to the Analysis of Closed-Loop Systems by M.B. Priestley
14. The Bispectral Analysis of Nonlinear Stationary Time Series with Reference to Bilinear Time-Series Models by T. Subba Rao
15. Frequency-Domain Analysis of Multidimensional Time-Series Data by E.A. Robinson

16. Review of Various Approaches to Power Spectrum Estimation by P.M. Robinson
17. Cumulants and Cumulant Spectra by M. Rosenblatt
18. Replicated Time-Series Regression: An Approach to Signal Estimation and Detection by R.H. Shumway
19. Computer Programming of Spectrum Estimation by T. Thrall
20. Likelihood Ratio Tests on Covariance Matrices and Mean Vectors of Complex Multivariate Normal Populations and their Applications in Time Series by P.R. Krishnaiah, J.C. Lee and T.C. Chang

Volume 4. Nonparametric Methods
Edited by P.R. Krishnaiah and P.K. Sen
1984 xx + 968 pp.

1. Randomization Procedures by C.B. Bell and P.K. Sen
2. Univariate and Multivariate Multisample Location and Scale Tests by V.P. Bhapkar
3. Hypothesis of Symmetry by M. Hušková
4. Measures of Dependence by K. Joag-Dev
5. Tests of Randomness against Trend or Serial Correlations by G.K. Bhattacharyya
6. Combination of Independent Tests by J.L. Folks
7. Combinatorics by L. Takács
8. Rank Statistics and Limit Theorems by M. Ghosh
9. Asymptotic Comparison of Tests – A Review by K. Singh
10. Nonparametric Methods in Two-Way Layouts by D. Quade
11. Rank Tests in Linear Models by J.N. Adichie
12. On the Use of Rank Tests and Estimates in the Linear Model by J.C. Aubuchon and T.P. Hettmansperger
13. Nonparametric Preliminary Test Inference by A.K.Md.E. Saleh and P.K. Sen
14. Paired Comparisons: Some Basic Procedures and Examples by R.A. Bradley
15. Restricted Alternatives by S.K. Chatterjee
16. Adaptive Methods by M. Hušková
17. Order Statistics by J. Galambos
18. Induced Order Statistics: Theory and Applications by P.K. Bhattacharya
19. Empirical Distribution Function by F. Csáki
20. Invariance Principles for Empirical Processes by M. Csörgő
21. M-, L- and R-estimators by J. Jurečková
22. Nonparametric Sequential Estimation by P.K. Sen
23. Stochastic Approximation by V. Dupač
24. Density Estimation by P. Révész
25. Censored Data by A.P. Basu
26. Tests for Exponentiality by K.A. Doksum and B.S. Yandell
27. Nonparametric Concepts and Methods in Reliability by M. Hollander and F. Proschan
28. Sequential Nonparametric Tests by U. Müller-Funk
29. Nonparametric Procedures for some Miscellaneous Problems by P.K. Sen
30. Minimum Distance Procedures by R. Beran

Volume 5. Time Series in the Time Domain
Edited by E.J. Hannan, P.R. Krishnaiah and M.M. Rao
1985 xiv + 490 pp.

Volume 6. Sampling
Edited by P.R. Krishnaiah and C.R. Rao
1988 xvi + 594 pp.

Volume 7. Quality Control and Reliability

Edited by P.R. Krishnaiah and C.R. Rao

1988 xiv + 503 pp.

Volume 8. Statistical Methods in Biological and Medical Sciences
Edited by C.R. Rao and R. Chakraborty
1991 xvi + 554 pp.

Volume 9. Computational Statistics

Edited by C.R. Rao

1993 xix + 1045 pp.

Volume 10. Signal Processing and its Applications
Edited by N.K. Bose and C.R. Rao
1993 xvii + 992 pp.

Volume 11. Econometrics

Edited by G.S. Maddala, C.R. Rao and H.D. Vinod

1993 xx + 783 pp.

Volume 12. Environmental Statistics
Edited by G.P. Patil and C.R. Rao
1994 xix + 927 pp.

Volume 13. Design and Analysis of Experiments
Edited by S. Ghosh and C.R. Rao
1996 xviii + 1230 pp.

Volume 14. Statistical Methods in Finance
Edited by G.S. Maddala and C.R. Rao
1996 xvi + 733 pp.

Volume 15. Robust Inference
Edited by G.S. Maddala and C.R. Rao
1997 xviii + 698 pp.

5. Introduction to Positive-Breakdown Methods by P.J. Rousseeuw
6. Outlier Identification and Robust Methods by U. Gather and C. Becker
7. Rank-Based Analysis of Linear Models by T.P. Hettmansperger, J.W. McKean and S.J. Sheather
8. Rank Tests for Linear Models by R. Koenker
9. Some Extensions in the Robust Estimation of Parameters of Exponential and Double Exponential Distributions in the Presence of Multiple Outliers by A. Childs and N. Balakrishnan
10. Outliers, Unit Roots and Robust Estimation of Nonstationary Time Series by G.S. Maddala and Y. Yin
11. Autocorrelation-Robust Inference by P.M. Robinson and C. Velasco
12. A Practitioner's Guide to Robust Covariance Matrix Estimation by W.J. den Haan and A. Levin
13. Approaches to the Robust Estimation of Mixed Models by A.H. Welsh and A.M. Richardson
14. Nonparametric Maximum Likelihood Methods by S.R. Cosslett
15. A Guide to Censored Quantile Regressions by B. Fitzenberger
16. What Can Be Learned About Population Parameters When the Data Are Contaminated by J.L. Horowitz and C.F. Manski
17. Asymptotic Representations and Interrelations of Robust Estimators and Their Applications by J. Jurecková and P.K. Sen
18. Small Sample Asymptotics: Applications in Robustness by C.A. Field and M.A. Tingley
19. On the Fundamentals of Data Robustness by G. Maguluri and K. Singh
20. Statistical Analysis With Incomplete Data: A Selective Review by M.G. Akritas and M.P. La Valley
21. On Contamination Level and Sensitivity of Robust Tests by J.Á. Visšek
22. Finite Sample Robustness of Tests: An Overview by T. Kariya and P. Kim
23. Future Directions by G.S. Maddala and C.R. Rao

Volume 16. Order Statistics – Theory and Methods

Edited by N. Balakrishnan and C.R. Rao

1997 xix + 688 pp.

1. Order Statistics: An Introduction by N. Balakrishnan and C.R. Rao
2. Order Statistics: A Historical Perspective by H. Leon Harter and N. Balakrishnan
3. Computer Simulation of Order Statistics by Pandu R. Tadikamalla and N. Balakrishnan
4. Lorenz Ordering of Order Statistics and Record Values by Barry C. Arnold and Jose A. Villasenor
5. Stochastic Ordering of Order Statistics by Philip J. Boland, Moshe Shaked and J. George Shanthikumar
6. Bounds for Expectations of L-Estimates by T. Rychlik
7. Recurrence Relations and Identities for Moments of Order Statistics by N. Balakrishnan and K.S. Sultan

8. Recent Approaches to Characterizations Based on Order Statistics and Record Values by C.R. Rao and D.N. Shanbhag
9. Characterizations of Distributions via Identically Distributed Functions of Order Statistics by Ursula Gather, Udo Kamps and Nicole Schweitzer
10. Characterizations of Distributions by Recurrence Relations and Identities for Moments of Order Statistics by Udo Kamps
11. Univariate Extreme Value Theory and Applications by Janos Galambos
12. Order Statistics: Asymptotics in Applications by Pranab Kumar Sen
13. Zero-One Laws for Large Order Statistics by R.J. Tomkins and Hong Wang
14. Some Exact Properties of Cook's D_1 by D.R. Jensen and D.E. Ramirez
15. Generalized Recurrence Relations for Moments of Order Statistics from Non-Identical Pareto and Truncated Pareto Random Variables with Applications to Robustness by Aaron Childs and N. Balakrishnan
16. A Semiparametric Bootstrap for Simulating Extreme Order Statistics by Robert L. Strawderman and Daniel Zelterman
17. Approximations to Distributions of Sample Quantiles by Chunsheng Ma and John Robinson
18. Concomitants of Order Statistics by H.A. David and H.N. Nagaraja
19. A Record of Records by Valery B. Nevzorov and N. Balakrishnan
20. Weighted Sequential Empirical Type Processes with Applications to Change-Point Problems by Barbara Szyszkowicz
21. Sequential Quantile and Bahadur–Kiefer Processes by Miklós Csörgő and Barbara Szyszkowicz

Volume 17. Order Statistics: Applications
Edited by N. Balakrishnan and C.R. Rao
1998 xviii + 712 pp.

1. Order Statistics in Exponential Distribution by Asit P. Basu and Bahadur Singh
2. Higher Order Moments of Order Statistics from Exponential and Right-truncated Exponential Distributions and Applications to Life-testing Problems by N. Balakrishnan and Shanti S. Gupta
3. Log-gamma Order Statistics and Linear Estimation of Parameters by N. Balakrishnan and P.S. Chan
4. Recurrence Relations for Single and Product Moments of Order Statistics from a Generalized Logistic Distribution with Applications to Inference and Generalizations to Double Truncation by N. Balakrishnan and Rita Aggarwala
5. Order Statistics from the Type III Generalized Logistic Distribution and Applications by N. Balakrishnan and S.K. Lee
6. Estimation of Scale Parameter Based on a Fixed Set of Order Statistics by Sanat K. Sarkar and Wenjin Wang
7. Optimal Linear Inference Using Selected Order Statistics in Location-Scale Models by M. Masoom Ali and Dale Umbach
8. L-Estimation by J.R.M. Hosking
9. On Some L-estimation in Linear Regression Models by Soroush Alimoradi and A.K.Md. Ehsanes Saleh

Volume 18. Bioenvironmental and Public Health Statistics

Edited by P.K. Sen and C.R. Rao

2000 xxiv + 1105 pp.

Volume 19. Stochastic Processes: Theory and Methods
Edited by D.N. Shanbhag and C.R. Rao
2001 xiv + 967 pp.

Volume 20. Advances in Reliability
Edited by N. Balakrishnan and C.R. Rao
2001 xxii + 860 pp.

Volume 21. Stochastic Processes: Modelling and Simulation
Edited by D.N. Shanbhag and C.R. Rao
2003 xxviii + 1002 pp.

Volume 22. Statistics in Industry
Edited by R. Khattree and C.R. Rao
2003 xxi + 1150 pp.

Volume 23. Advances in Survival Analysis
Edited by N. Balakrishnan and C.R. Rao
2003 xxv + 795 pp.

Volume 24. Data Mining and Data Visualization
Edited by C.R. Rao, E.J. Wegman and J.L. Solka
2005 xiv + 643 pp.

7. Canonical Variate Analysis and Related Methods for Reduction of Dimensionality and Graphical Representation by C.R. Rao
8. Pattern Recognition by D.J. Hand
9. Multidimensional Density Estimation by D.W. Scott and S.R. Sain
10. Multivariate Outlier Detection and Robustness by M. Hubert, P.J. Rousseeuw and S. Van Aelst
11. Classification and Regression Trees, Bagging, and Boosting by C.D. Sutton
12. Fast Algorithms for Classification Using Class Cover Catch Digraphs by D.J. Marchette, E.J. Wegman and C.E. Priebe
13. On Genetic Algorithms and their Applications by Y.H. Said
14. Computational Methods for High-Dimensional Rotations in Data Visualization by A. Buja, D. Cook, D. Asimov and C. Hurley
15. Some Recent Graphics Templates and Software for Showing Statistical Summaries by D.B. Carr
16. Interactive Statistical Graphics: the Paradigm of Linked Views by A. Wilhelm
17. Data Visualization and Virtual Reality by J.X. Chen

Volume 25. Bayesian Thinking: Modeling and Computation
Edited by D.K. Dey and C.R. Rao
2005 xx + 1041 pp.

1. Bayesian Inference for Causal Effects by D.B. Rubin
2. Reference Analysis by J.M. Bernardo
3. Probability Matching Priors by G.S. Datta and T.J. Sweeting
4. Model Selection and Hypothesis Testing based on Objective Probabilities and Bayes Factors by L.R. Pericchi
5. Role of P-values and other Measures of Evidence in Bayesian Analysis by J. Ghosh, S. Purkayastha and T. Samanta
6. Bayesian Model Checking and Model Diagnostics by H.S. Stern and S. Sinharay
7. The Elimination of Nuisance Parameters by B. Liseo
8. Bayesian Estimation of Multivariate Location Parameters by A.C. Brandwein and W.E. Strawderman
9. Bayesian Nonparametric Modeling and Data Analysis: An Introduction by T.E. Hanson, A.J. Branscum and W.O. Johnson
10. Some Bayesian Nonparametric Models by P. Damien
11. Bayesian Modeling in the Wavelet Domain by F. Ruggeri and B. Vidakovic
12. Bayesian Nonparametric Inference by S. Walker
13. Bayesian Methods for Function Estimation by N. Choudhuri, S. Ghosal and A. Roy
14. MCMC Methods to Estimate Bayesian Parametric Models by A. Mira
15. Bayesian Computation: From Posterior Densities to Bayes Factors, Marginal Likelihoods, and Posterior Model Probabilities by M.-H. Chen
16. Bayesian Modelling and Inference on Mixtures of Distributions by J.-M. Marin, K. Mengersen and C.P Robert
17. Simulation Based Optimal Design by P. Müller

Volume 26. Psychometrics
Edited by C.R. Rao and S. Sinharay
2007 xx + 1169 pp.

Volume 27. Epidemiology and Medical Statistics
Edited by C.R. Rao, J.P. Miller, and D.C.Rao
2009 xviii + 812 pp.

22. Power: Establishing the Optimum Sample Size by Richard A. Zeller and Yan Yan
23. Statistical Learning in Medical Data Analysis by Grace Wahba
24. Evidence Based Medicine and Medical Decision Making by Dan Mayer, MD
25. Estimation of Marginal Regression Models with Multiple Source Predictors by Heather J. Litman, Nicholas J. Horton, Bernardo Hernández and Nan M. Laird
26. Difference Equations with Public Health Applications by Asha Seth Kapadia and Lemuel A. Moyé
27. The Bayesian Approach to Experimental Data Analysis by Bruno Lecoutre

Volume 29A. Sample Surveys: Design, Methods and Applications
Edited by Danny Pfeffermann and C. R. Rao
2009 xxiv + 698 pp.

1. Introduction to Survey Sampling by Ken Brewer and Timothy G. Gregoire
2. Sampling with Unequal Probabilities by Yves G. Berger and Yves Tillé
3. Two-Phase Sampling by Jason C. Legg and Wayne A. Fuller
4. Multiple-Frame Surveys by Sharon L. Lohr
5. Designs for Surveys over Time by Graham Kalton
6. Sampling of Rare Populations by Mary C. Christman
7. Design, Conduct, and Analysis of Random-Digit Dialing Surveys by Kirk Wolter, Sadeq Chowdhury and Jenny Kelly
8. Nonresponse and Weighting by J. Michael Brick and Jill M. Montaquila
9. Statistical Data Editing by Ton De Waal
10. Imputation and Inference in the Presence of Missing Data by David Haziza
11. Dealing with Outliers in Survey Data by Jean-François Beaumont and Louis-Paul Rivest
12. Measurement Errors in Sample Surveys by Paul Biemer
13. Computer Software for Sample Surveys by Jelke Bethlehem
14. Record Linkage by William E. Winkler
15. Statistical Disclosure Control for Survey Data by Chris Skinner
16. Sampling and Estimation in Household Surveys by Jack G. Gambino and Pedro Luis do Nascimento Silva
17. Sampling and Estimation in Business Surveys by Michael A. Hidiroglou and Pierre Lavallée
18. Sampling, Data Collection, and Estimation in Agricultural Surveys by Sarah M. Nusser and Carol C. House
19. Sampling and Inference in Environmental Surveys by David A. Marker and Don L. Stevens Jr.
20. Survey Sampling Methods in Marketing Research: A Review of Telephone, Mall Intercept, Panel, and Web Surveys by Raja Velu and Gurramkonda M. Naidu
21. Sample Surveys and Censuses by Ronit Nirel and Hagit Glickman
22. Opinion and Election Polls by Kathleen A. Frankovic, Costas Panagopoulos and Robert Y. Shapiro

Volume 29B. Sample Surveys: Inference and Analysis
Edited by Danny Pfeffermann and C. R. Rao
2009 xxiv + 642 pp.

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank