CHAPTER 6

PARAMETER ESTIMATION

6.1 INTRODUCTION

In Chapter 5, a range of informative graphical methods are presented for identifying the parameters to include in an ARMA model for fitting to a given nonseasonal time series. As explained in that chapter, prior to deciding upon the form of the ARMA model, the data may be transformed using the Box-Cox transformation in [3.4.30] in order to alleviate problems with nonnormality and/or changing variance. Additionally, the differencing given in [4.3.3] may be required for removing nonstationarity. Whatever the case, at the identification stage, one must select one or more appropriate ARMA models from [3.4.4] or [4.3.4] for fitting to either the series as given or some modified version thereof.

As shown in Figure III.1, subsequent to identifying one or more tentative models for fitting to a particular series, one must obtain estimates for the parameters in the models. The main objectives of Chapter 6 are to present procedures for *estimating the parameters in ARMA models* and to explain how an automatic selection criterion such as the *Akaike information criterion* (Akaike, 1974) can be employed for choosing the best overall model when more than one model is calibrated.

For an identified ARMA model in [3.4.4] or [4.3.4], the following parameters must be estimated using the available data:

1. mean of the series,
2. AR parameters,
3. MA parameters,
4. innovation series,
5. variance of the innovations.

Because one often knows a priori the best type of *Box-Cox transformation* to use with a given kind of time series such as annual riverflows, one can first fix λ in [3.4.30] at a specified value before estimating the model parameters mentioned above. If λ is not known, it is possible to estimate λ along with the other model parameters. However, this requires a significant increase in the amount of computer time needed to estimate all the model parameters. Finally, one should keep in mind that the integer value for the *differencing parameter d* contained in ARIMA(p,d,q) models in Chapter 4 is selected using identification methods (see Sections 5.3.3 and 5.3.4). If differencing is used, often one may wish to fix the mean of the differenced series at zero and not estimate it (see discussion in Section 4.3.1). When d is allowed to take on real values to form the fractional ARMA models described in Chapter 11, one must estimate the value of d.

A given time series is just one possible realization or set of measurements from the phenomenon that generated it (see discussion in Sections 2.2 and 2.3). Because a time series contains only partial information about the phenomenon under study, the true or population values of the parameters of a model fitted to the series are not known. Consequently, there is *uncertainty* about the estimation of the model parameters. As explained in Section 6.2.3 and

Appendix A6.2, the uncertainty for a specified parameter estimate is quantified by what is called the *standard error* (SE) of the estimate.

Estimation theory was initiated by the great German mathematician Karl Friederich Gauss who developed the method of least squares for solving practical problems. Since the time of Gauss, well known researchers such as Sir R.A. Fisher, Norbert Wiener and R.E. Kalman, have developed an impressive array of estimation procedures and associated algorithms. These general approaches from estimation theory have been formulated for use with specific families of statistical models. For example, in this chapter the method of *maximum likelihood* is described and used for estimating the parameters of ARMA models. In Section 3.2.2, the Yule-Walker equations given in [3.2.11] can be employed for obtaining what are called *moment estimates* for AR models.

A great number of textbooks and research papers about estimation theory are available. Mendel's (1987) book, for example, covers a wide variety of estimation techniques including least squares, maximum likelihood and the Kalman filter (Kalman, 1960) approaches. A research paper by Norden (1972, 1973) presents a survey of maximum likelihood estimation which was originally developed by Fisher (1922, 1925). The monograph of Edwards (1972) also deals with the maximum likelihood approach to estimation. Most textbooks, in statistics, such as the ones by Kempthorne and Folks (1971) and Cox and Hinkley (1974) contain large sections dealing with estimation. In addition, statistical encyclopediae (Kotz and Johnson, 1988; Kruskal and Tanur, 1978; Kendall and Buckland, 1971) and handbooks (Sachs, 1984) have good explanations about estimation procedures.

Because of many attractive theoretical properties, maximum likelihood estimation is the most popular general approach to parameter estimation. In the next section, some of these properties are pointed out and maximum likelihood estimation for calibrating ARMA models is described. Subsequent to this, it is explained how the Akaike information criterion (Akaike, 1974) can be used to select the overall best model when more than one model is fitted to a specified time series. Practical applications are used for illustrating how estimation is carried out in practice and the Akaike information criterion can be used for model discrimination.

6.2 MAXIMUM LIKELIHOOD ESTIMATION

6.2.1 Introduction

The *probability distribution function* (pdf) of a set of random variables is written as a function of these variables and certain given parameters. For example, for the case of a single random variable following a normal distribution, the pdf is a function of this random variable and the parameters in the pdf are the mean and variance. When the actual values of the mean and variance are known for the normally distributed random variable, one can calculate the probability that the random variable takes on a value within a specified range by integrating the pdf over this range. On the other hand, if the measurements for a random variable are substituted into the pdf and the pdf is then considered as a function of the parameters that have not been estimated, the likelihood function is created. In other words, the *likelihood function* is essentially the probability of the actual data as a function of the parameters.

To be more specific, suppose that one is dealing with the sequence of observations in [4.3.3] which consists of n values represented by the vector $\mathbf{w}' = (w_1, w_2, \ldots, w_n)$. The sample of n observations, $\mathbf{w}$, can be associated with an n-dimensional random variable having a known pdf, $p(\mathbf{w}|\beta)$, which depends on a vector of unknown parameters β. For the case of the ARMA model in [3.4.4] or the ARIMA model in [4.3.4], the parameters contained in β are the p AR parameters $\phi = (\phi_1, \phi_2, \ldots, \phi_p)$, q MA parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$, and the variance, σ_a^2, of the innovations. Hence, $\beta = (\phi, \theta, \sigma_a^2)$.

In advance of having the data, the pdf given by $p(\mathbf{w}|\beta)$ associates a density with a possible realization of $\mathbf{w}$, for fixed β. When the observations are available, one would like to find out values of β which could have produced the set of time series entries, $\mathbf{w}$. To accomplish this, one substitutes the data, $\mathbf{w}$, into the pdf and considers β as the variable in order to produce the likelihood function $L(\beta|\mathbf{w})$. Because of the way it is defined, the likelihood function has the same form as $p(\mathbf{w}|\beta)$. However, in the likelihood function the set of observations, $\mathbf{w}$, is fixed and the parameters contained in β are variable.

Because the relative value of the likelihood function, $L(\beta|\mathbf{w})$, is of main interest, the likelihood function often contains an arbitrary multiplicative constant. For simplifying calculations, it is usually more convenient to use the natural logarithm of the likelihood function given by $\ln L(\beta|\mathbf{w}) = l(\beta|\mathbf{w})$, which possesses an arbitrary additive constant. This function is commonly referred to as the *log likelihood function.*

In maximum likelihood estimation, one wants to determine the values of the parameters contained in β that maximize the likelihood function or, equivalently, the log likelihood function. These estimates are called *maximum likelihood estimates (MLE's).*

One approach to finding the maximum value for a given function is to set the first order partial derivative with respect to each variable parameter equal to zero and then to solve these equations to find the values of the variables which maximize the function. Because the likelihood function for ARMA models is quite complicated, this simple approach cannot be used. Consequently, after defining the likelihood function for ARMA models in Section 6.2.3 and Appendix A6.1, some useful optimization algorithms are recommended for optimizing the likelihood function.

The second order partial derivative of the likelihood or log likelihood function with respect to each of the variable parameters reflects the rate of change of the slope or, in other words, the "spread" of the function. Accordingly, these second order derivatives, which are contained in the *information matrix* defined in Appendix A6.2, are used to determine approximate *standard errors (SE's)* for the MLE's. However, before going into the details of the likelihood function and the associated information matrix, the motivations for using maximum likelihood estimation are explained next.

6.2.2 Properties of Maximum Likelihood Estimators

Likelihood Principle

Prior to describing some of the attractive properties of maximum likelihood estimation, consider first an important characteristic of the likelihood function. One main reason why the likelihood function is of such great import in estimation theory is because of what is called the

likelihood principle summarized below.

Likelihood Principle: Assuming that the underlying model is correct, all the information that the data can provide about the model parameters is contained in the likelihood function. All other aspects of the data are irrelevant with respect to characterizing the model parameters (Fisher, 1956; Barnard, 1949; Birnhaum, 1962).

The likelihood principle is in consonance with the Bayesian approach to statistics. This is because the likelihood function is a component in the posterior distribution of the parameters coming from the data.

As noted in the previous section, when the likelihood function, or equivalently, the log likelihood function is maximized, one obtains MLE's for the model parameters. The general mathematical expression which defines how one obtains MLE's for any set of data for a given family of models, is called the *maximum likelihood estimator*. In Appendix A6.1, for example, a maximum likelihood estimator is presented for calculating MLE's for the parameters of an ARMA model fitted to a given time series.

In general, most maximum likelihood estimators possess some fundamental statistical properties which, in turn, have led to the widespread development, acceptance and application of these estimators. To characterize estimators, Fisher (1925) introduced the concepts of consistency and large-sample efficiency. Although these concepts are defined in terms of large samples, estimators having these characteristics are usually well suited for use in practical applications. Because maximum likelihood estimators usually satisfy these concepts, Fisher and many other statisticians have advocated their employment for application purposes. The maximum likelihood estimators referred to in Section 6.2.3 and the one described in Appendix A6.1 are consistent and efficient. These two concepts are now briefly summarized. For detailed mathematical definitions of the concepts, the reader can refer to the references cited in this section as well as statistical encyclopediae, handbooks and standard textbooks.

Consistency

A *consistent estimator* is one which converges in probability as the sample size increases to the true value of the parameter or parameters being estimated. More specifically, let $\hat{\eta}$ be the estimate of a model parameter η using a given estimator for a sample size of n. The estimate $\hat{\eta}$ of η, or equivalently its estimator, is consistent if

$$\lim_{n \to \infty} P[|\hat{\eta} - \eta| > \epsilon] = 0 \qquad [6.2.1]$$

where P stands for probability, and ϵ is any positive number which can, of course, be very close to zero.

In practice, one would like to have an estimator which produces estimates that converge to the true values of the model parameters as the sample size increases. Although exceptions can be found, most maximum likelihood estimators are consistent.

Efficiency

Suppose that two estimators are consistent. Let $\hat{\eta}_1$ and $\hat{\eta}_2$ denote the two consistent estimators or estimates for a model parameter η where the sample size is n. The *asymptotic relative efficiency (ARE)* of $\hat{\eta}_1$ with respect to $\hat{\eta}_2$ is:

$$ARE = \lim_{n \to \infty} \frac{var\ \hat{\eta}_1}{var\ \hat{\eta}_2} \qquad [6.2.2]$$

If the above ratio is less (more) than one, the estimator $\hat{\eta}_1$ is asymptotically more (less) efficient than $\hat{\eta}_2$ for estimating η. When the limit is equal to one, the estimators are equally efficient. The asymptotic relative efficiency is the limiting value of the *relative efficiency (RE)* given by:

$$RE = \frac{var\hat{\eta}_1}{var\hat{\eta}_2} \qquad [6.2.3]$$

For maximum likelihood estimators or MLE's, the variance of the MLE of a model parameter possesses minimum asymptotic variance and is asymptotically normally distributed when consistency and other conditions are satisfied (Cramer, 1946; Rao, 1973). Therefore, when investigating the properties of a given estimator, it is informative to compare it to its MLE counterpart. Suppose that $\hat{\eta}_1$ is the maximum likelihood estimator for a model parameter η and $\hat{\eta}_2$ is another estimator in [6.2.2]. Because the maximum likelihood estimator possesses minimum variance $0 \leq ARE \leq 1$. Furthermore, the ratio is referred to as the *first order asymptotic efficiency* of $\hat{\eta}_2$ with respect to the maximum likelihood estimator $\hat{\eta}_1$. If the first order efficiency is less than unity, the estimator $\hat{\eta}_2$ is less efficient than the maximum likelihood estimator $\hat{\eta}_1$ for large samples. However, when the first order efficiency is equal to one, the ratio in [6.2.2] cannot distinguish between the two estimators. One must then examine what is called second order efficiency (Rao, 1961, 1962) in order to select the most efficient estimator. *Second order efficiency* is concerned with the speed of convergence of the ratio in [6.2.2] and usually requires rather complicated expressions in order to be properly defined. Whatever the case, the maximum likelihood estimator is the only known estimator that possesses second order efficiency.

Gaussian Efficiency: In the definitions of the ARMA family of models in [3.4.4] and the ARIMA class of models in [4.3.4], the innovation series represented by a_t is assumed to be identically and independently distributed with a mean of zero and variance of σ_a^2 [i.e. IID$(0,\sigma_a^2)$]. To allow one to derive the likelihood function for these models, one must specify a distribution for the innovations. In practice, the a_t's are assumed to be normally independently distributed with a mean of zero and variance of σ_a^2 [i.e. NID$(0,\sigma_a^2)$]. The likelihood function for the Gaussian or normal case is discussed in Section 6.2.3 and presented in detail in Appendix A6.1. By determining the values of the model parameters which maximize the value of the likelihood or log likelihood function, one determines MLE's for the parameters. As explained in Appendix A6.2, the *covariance matrix* is obtained as the inverse of the *information matrix* and the entries along the diagonal give the variance of the estimates for the corresponding model parameters. The square root of these variances are called the SE's of estimation for the model parameters. Because the maximum likelihood procedure is used to obtain the parameter estimates, these SE's or, equivalently, the variances, possess Fisherian efficiency and, therefore, are

the smallest values that can be obtained in large samples.

If the innovations are not normally distributed, one can use the same technique as for the NID case to obtain estimates for the model parameters. Even though the innovations do not follow a normal distribution, these estimates are called *Gaussian estimates* because the maximum likelihood estimator for NID innovations is used to calculate the estimates. It can be shown theoretically that the large sample covariance matrix for Gaussian estimates is the same as that for the situations for which the innovations are NID. This robustness property of maximum likelihood estimation under the normality assumption is referred to as *Gaussian efficiency* (Whittle, 1961; Hannan, 1970, pp. 377-383). However, the reader should keep in mind that even though the Gaussian estimates possess Gaussian efficiency, they are not Fisherian efficient (i.e. have minimum variances for the estimates) because the innovations do not follow a normal distribution.

Li and McLeod (1988) show how maximum likelihood may be used to fit ARMA models when the innovations, a_t, are non-Gaussian. For example, when the a_t are log-normal or gamma distributed, improved estimates of the parameters can be obtained by using maximum likelihood estimation.

6.2.3 Maximum Likelihood Estimators

A given nonseasonal series, z_t, may first be transformed using the Box-Cox transformation in [3.4.30] in order to make the series approximately normally distributed. Subsequent to a power transformation, the series can be differenced as in [4.3.3] just enough times to remove any nonstationarity. One then ends up with a stationary series w_t, $t = 1,2,\ldots,n$, which follows a normal distribution. By employing the identification procedures of Section 5.3, one can decide upon an appropriate ARMA(p,q) model to fit to the w_t series. Of course, if no Box-Cox transformation or differencing are needed, the w_t series is simply the original z_t observations.

Assuming that the innovations in [4.3.4] or [3.4.4] are NID, which also implies that the w_t or z_t sequences follow a normal distribution, one can derive the likelihood function for an ARMA model. By employing a suitable optimization algorithm to maximize the likelihood or log likelihood function with respect to the ARMA model parameters, one should theoretically be able to obtain MLE's for the parameters. However, the likelihood function is a fairly complicated expression and flexible algorithms are needed in order to make it computationally possible within a reasonable amount of time to maximize the likelihood function in order to find the MLE's. As a result, researchers have suggested saving computational time by maximizing approximations to the likelihood function to calculate *approximate MLE's* for the model parameters. As the sample size increases, the approximate MLE's approach closer and closer to the true MLE's. Box and Jenkins (1976, Ch. 7), for example, have put forward two approximate maximum likelihood procedures for ARMA models which are called the conditional and the unconditional or iterated methods. Generally speaking, their approaches do not work as well for ARMA models containing MA parameters and for time series that are fairly short (McLeod, 1977).

McLeod (1977) derives an approximate maximum likelihood procedure which is almost exact. His technique is referred to as the *modified sum of squares algorithm*. Besides providing parameter estimates that are very close to the true or exact MLE's, the approach is very efficient

computationally and, therefore, requires relatively little computer time. Moreover, it works well with models containing MA parameters and series having relatively few observations.

More recently, a number of authors have developed *exact maximum likelihood estimators* for use with ARMA models. These exact techniques include contributions by:

1. Newbold (1974),
2. Ansley (1979),
3. Ljung and Box (1979), and
4. Mélard (1984) who uses a Kalman filter approach to maximum likelihood estimation.

As just noted, McLeod's (1977) estimation technique for ARMA models is computationally efficient and produces estimates that are almost exact MLE's. Furthermore, the procedure has been extended for use with seasonal ARMA models (McLeod and Sales, 1983). Accordingly, this flexible algorithm is recommended for use in practical ARMA modelling and is outlined in Appendix A6.1. The McLeod-Hipel time series package referred to in Section 1.7 contains the estimation algorithm of Appendix A6.1 as well as other approximate and exact maximum likelihood estimators.

In addition to possessing desirable statistical properties, maximum likelihood estimation is computationally convenient. This is because a range of useful and powerful optimization techniques are available to maximize or minimize a function such as the likelihood or log likelihood function with respect to the model parameters. Some of the optimization algorithms that have been extensively utilized in practical applications include:

1. Gauss linearization (Draper and Smith, 1980),
2. steepest descent (Draper and Smith, 1980),
3. Marquardt algorithm which is a combination of the above two algorithms (Marquardt, 1963),
4. conjugate directions (Powell, 1964, 1965).
5. Davidon's Algorithm (Davidon, 1968) for which a FORTRAN subroutine is provided by Ishiguro and Akaike (1989) for log likelihood maximization.

For the applications given in Section 6.4 and other chapters in this textbook conjugate directions is used in conjunction with the estimation procedure of Appendix A6.1 to obtain MLE's for the ARMA model parameters. For an explanation of the variety of optimization methods, the reader can refer to textbooks such as those by Luenberger (1984), Gill et al. (1981) and VanderPlaats (1984).

To obtain estimates for the parameters in an ARMA model, a time series of observations is used with an appropriate maximum likelihood estimator. Because this time series is only a finite sample realization of the phenomenon generating the series, the MLE for a given parameter is not the population value. The SE or standard derivation of the estimate is used to reflect the *uncertainty* contained in the estimate. In Appendix A6.2, it is explained how the SE's for the parameter estimates are defined. More specifically, the variance-covariance matrix of the parameter estimates is the inverse of what is called the information matrix. The square roots of the diagonal entries in the variance-covariance matrix provide the estimates of the SE's for the estimated model parameters. Furthermore, because it is known that MLE's are asymptotically

normally distributed, one can obtain 95% confidence limits for a given parameter estimate. If for example, zero were contained within the interval formed by a parameter estimate ± 1.96SE, one could argue that the parameter estimate is not significantly different from zero and perhaps the parameter should be left out of the model.

6.3 MODEL DISCRIMINATION USING THE AKAIKE INFORMATION CRITERION

6.3.1 Introduction

As noted in Chapter 5, the practitioner is usually confronted with the problem of choosing the most appropriate model for fitting to a given data set from a large number of available models. Consequently, *model discrimination* procedures are required and some possible selection methods are listed in Section 5.2.3. The identification methods in Section 5.3 constitute graphical and tabular techniques that can assist in deciding upon which model to choose. However, these methods require some skill when being used in applications since the modeller must be cognizant of the properties of the various types of identification graphs in order to ascertain which parameters should be included in the model. To increase the speed, flexibility, accuracy and simplicity involved in choosing a model, the *Akaike Information Criterion (AIC)* (Akaike, 1974) has been found to be quite useful. The AIC was first suggested for use in hydrology by Hipel et al. (1977) and McLeod et al. (1977). Hipel (1981) explains in detail how the AIC can be used in geophysical model discrimination and provides references for its application to many different kinds of time series.

6.3.2 Definition of the Akaike Information Criterion

Based upon information theory, Akaike (1972a, 1973, 1974) developed the AIC which is defined as

$$AIC = -2\ln ML + 2k \qquad [6.3.1]$$

where ML denotes maximum likelihood, $\ln ML$ is the value of the maximized log likelihood function for a model fitted to a given data set, and k is the number of independently adjusted parameters within the model. A desirable attribute of the AIC is that the modelling principles described in Sections 1.3 and 5.2.4 are formally incorporated into the equation. The first term on the right hand side of [6.3.1] reflects the doctrine of *good statistical fit* while the second entry accounts for *model parsimony*. Because of the form of [6.3.1], when there are several available models for modelling a given time series, the model that possesses the minimum value of the AIC should be selected. This procedure is referred to by Akaike (1974) as *MAICE (minimum AIC estimation)*.

The original mathematical development for the AIC formula in [6.3.1] is given by Akaike (1973, 1974) while a summary of the derivation is presented by Ozaki (1977) and also Kitagawa (1979). Even though the entries in [6.3.1] reflect sound modelling principles, as noted by Akaike (1978) "... the only justification of its use will come from its performance in applications." The MAICE procedure has previously been successfully applied to a wide range of statistical problems. The method has been used to decide upon the order of an ARMA model to fit to a time series (Akaike, 1974; Hipel et al., 1977; McLeod et al., 1977; Ozaki, 1977), to ascertain the type of nonstationary ARIMA model to describe a time series (Ozaki, 1977), to determine the order of an AR model (Akaike, 1978; Akaike, 1979; Shibata, 1976), to select the order of a

Markov chain (Tong, 1975), to decide upon the order of a polynomial regression (Akaike, 1972b; Tanabe, 1974), to determine the number of factors needed in a factor analysis (Akaike, 1971), to assist in robot data screening (Akaike, 1972b), to detect outliers in a data set (Kitagawa, 1979), to analyze cross classified data (Sakamoto and Akaike, 1977), and to assist in canonical correlation analysis of time series (Akaike, 1976). The AIC can be employed to select the most suitable model when more than one family of models are being considered and McLeod and Hipel (1978) used the MAICE procedure to determine whether an ARMA or Fractional Gaussian noise model should be utilized to model a given annual hydrological time series (see Section 10.4). The AIC can be employed to select the best model from the families of seasonal models discussed in Part VI, and to choose the most appropriate intervention model (see Chapter 19). In fact, the MAICE procedure can be used with all the models considered in this book (see Table 1.6.2) and the wide range of applications presented by Hipel (1981) confirm the versatility of this method for selecting the most appropriate model to fit to a time series.

6.3.3 The Akaike Information Criterion in Model Construction

Employment of the MAICE procedure reinforces and complements the identification, estimation and diagnostic stages of model constructions illustrated in Figure III.1 at the start of Part III. Figure 6.3.1 depicts how MAICE can be incorporated into the three stages of model development. Even though this chapter is concerned with nonseasonal ARMA and ARIMA models, the same general methodology can be employed no matter what types of time series models are being considered. For instance, the AIC model building procedure is recommended for use with the long memory, seasonal, transfer function-noise, intervention and multivariate models of Parts V to IX, respectively.

As shown by the flow chart in Figure 6.3.1, there are basically two approaches for employing the MAICE procedure in model construction. One method is to calculate the AIC for all possible models which are considered worthwhile for fitting to a given data set. For example, after specifying the Box-Cox parameter λ in [3.4.30] (often λ is set equal to unity if it is not known beforehand that a transformation is needed) maximum values for p, q and perhaps d may be set for ARIMA(p,d,q) models. The AIC can then be calculated for all possible combinations of p, d and q and the ARIMA model with the minimum AIC value is chosen. Although the selected model can usually be shown to adequately satisfy the important residual assumptions, as shown in Figure 6.3.1 it is always advisable to check for whiteness, normality, and homoscedasticity of the residuals using the methods in Sections 7.3 to 7.5, respectively. When the residuals are not white, other models should be considered by specifying a more flexible range for p, d and q. If the residuals do not possess constant variance and perhaps are not normally distributed then a suitable Box-Cox transformation may rectify the situation. To select the most suitable value of λ, a range of values of λ may be tried for the best ARIMA(p,d,q) model which was just chosen using the MAICE procedure. The value of λ which minimizes the AIC for the ARIMA model is then chosen. Another method is to obtain a MLE of λ for the best model and then to fix λ at this value if the exhaustive enumeration is repeated. Of course, λ could be estimated for all possible combinations of p, d and q in the exhaustive enumeration, but this would require a very large amount of computer usage.

If the diagnostic check stage is skipped and information from the identification and estimation stages is ignored when employing the exhaustive enumeration procedure with the AIC, it is possible that the best model may be missed. For example, in Section 6.4.3 it is shown that the

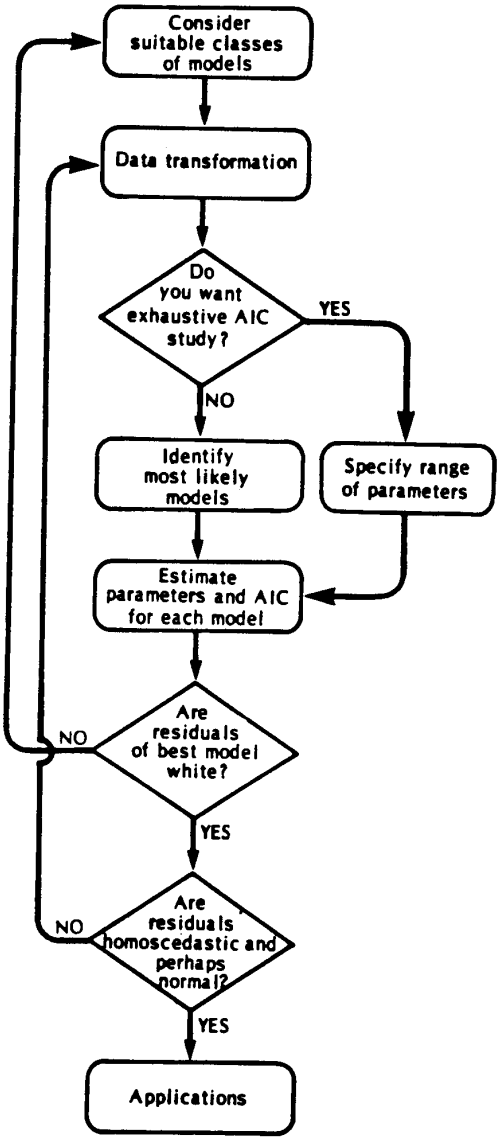

Figure 6.3.1. Model construction using MAICE.

most suitable type of ARMA model to fit to the annual sunspot series is an AR model of order 9 with the third to eighth AR parameter omitted from the model and the data transformed by a square root transformation. As explained in that section, if diagnostic testing had not been done and the SE's of the parameter estimates had not been considered, the most suitable model would not have been discovered. Besides the annual sunspot series, the MAICE procedure is used in Section 6.4.2 to decide upon the most appropriate ARMA model to fit to the average annual flows of the St. Lawrence River at Ogdensburg, New York.

An exhaustive AIC study may prove to be rather expensive due to the amount of computations. Consequently, as illustrated in Figure 6.3.1 an alternative approach is to only estimate the parameters and hence the AIC for a subset of models. For example, information from the identification stage (see Chapter 5) may indicate three tentative models to fit to the time series. The AIC is then only calculated for these three models and the model with the minimum AIC value is selected. If there are any problems with the residuals, appropriate action may be taken as shown in Figure 6.3.1. Otherwise, the chosen model can be employed in practical applications such as forecasting (see Chapter 8) or simulation (Chapter 9).

6.3.4 Plausibility

A question which is often asked by practitioners is how to interpret the relative differences in the values of the AIC for the various models which are fit to a specified data set. In fact, the different AIC values can be interpreted in a variety of manners. For example, if one model possesses an AIC value which is approximately $2k$ less than that of another model, this is analogous to the superior model having k less parameters than the other model. A lower AIC value can also be considered to be mainly due to a better statistical fit because of the first term on the right hand side of [6.3.1]. However, a lower AIC value is usually caused by both components of the formula in [6.3.1] and, therefore, an alternative approach for interpreting the differences in the AIC values between two models is to consider plausibility.

As shown by Akaike (1978), $\exp(-0.5AIC)$ is asymptotically a reasonable definition of the plausibility of a model specified by the parameters which are determined by the method of maximum likelihood. Consequently, the *plausibility* of model i versus model j can be calculated using

$$\text{Plausibility} = \exp[0.5(AIC_j - AIC_i)] \qquad [6.3.2]$$

where AIC_i is the value of the AIC for the ith model and AIC_j is the AIC value for the jth model. Table 6.3.1 displays some representative results for the plausibility of model i against model j where the jth model is assumed to have a lower AIC value than model i. As can be seen in Table 6.3.1, it is only the relative difference of the AIC values that is important and as these differences increase the plausibility decreases exponentially. Notice that when the AIC values differ by 6 the plausibility is only about 5%.

6.3.5 Akaike Information Criterion for ARMA and ARIMA Models

To determine the value of the AIC for an ARMA(p,q) model, both terms in [6.3.1] must be calculated separately. By optimizing the log likelihood function with respect to the model parameters (see Section 6.2.3 and Appendix A6.1), the value of the maximized log likelihood can be found for substitution into 6.3.1. The number of model parameters k is due to p AR parameters, q MA parameters, the variance of the model residuals, the Box-Cox exponent λ if it is included in the model, and the mean of the transformed series.

When considering a nonstationary series of length N, the data is differenced d times using [4.3.3] to produce a stationary series of length $n = N - d$. Because the differencing reduces the amount of information, this will certainly affect the first term on the right hand side of [6.3.1]. Hence, the AIC for an ARIMA model can be roughly calculated as

Table 6.3.1. Plausibility of model i versus model j.

$-(AIC_j - AIC_i)$	Plausibility
1	0.6065
2	0.3679
3	0.2313
4	0.1353
5	0.0821
6	0.0498
7	0.0302
8	0.0183
9	0.0111
10	0.0067
15	0.0006

$$AIC = \frac{N}{n}(-2\ln ML) + 2k \qquad [6.3.3]$$

where the value of the maximized log likelihood is obtained by optimizing the logarithm of [A6.1.5]. The total number of parameters k is the same as that for the ARMA model except when the mean of the differenced series is assumed to be zero and hence is not estimated, the number of parameters is decreased by one.

Another alternative for developing an AIC formula for an ARIMA model is to alter both components on the right hand side of [6.3.1]. As argued by Ozaki (1977), an increase in the number of data points contributes to decreasing the penalty due to the number of parameters. This effect can be incorporated into the AIC by writing the formula as

$$AIC = \frac{N}{n}(-2\ln ML + 2k) \qquad [6.3.4]$$

6.3.6 Other Automatic Selection Criteria

As shown by Figure 6.3.1 the MAICE procedure tends to "automate" model construction and to simplify model selection. In practice, it has been found that the MAICE methodology almost always chooses the same models which would be selected using more time consuming methods such as those presented in Section 5.3 and elsewhere. For example, when one model is a subset of another, a likelihood ratio test can be employed to determine if the model with more parameters is required for modelling a specified data set. However, as shown by McLeod et al. (1977), results from likelihood ratio tests usually confirm the conclusions reached using the MAICE procedure. An additional advantage of MAICE is that it is not necessary to select subjectively a significance level as is done with the likelihood ratio test.

The AIC is not the only *automatic selection criterion (ASC)* that can be used in model discrimination, although it is probably the most flexible and comprehensive of the methods which are presently available. For choosing the order of an AR model, Akaike (1969, 1970) developed the point estimation method called the final prediction error (FPE) technique (see Appendix A6.3 for a definition of the FPE and its relationship to the AIC). McLave (1975)

presented an algorithm to be used in subset autoregression for obtaining the best constrained AR model where model selection is based upon the FPE criterion (see Section 3.4.4 for a discussion of constrained models). In another paper, McLave (1978) compared the FPE technique and a sequential testing approach which he referred to as the "max χ^2 method" for choosing the constrained AR model. Other ASC's which can only be used for AR modelling include the technique devised by Anderson (1971), the "CAT" criterion of Parzen (1974), and the method of Hannan and Quinn (1979). The "D-statistic" of Gray et al. (1978) can be utilized for choosing the most appropriate nonseasonal ARMA(p,q) model, although the statistic has not been sufficiently developed for use in nonstationary and seasonal modelling. Mallows (1973) developed a statistic for use in model discrimination that is related to what he calls the C_p statistic. Based upon the characteristics of the sample ACF and the sample PACF (refer to Sections 5.3.4 and 5.3.5 for explanations of the sample ACF and PACF, respectively), Hill and Woodworth (1980) employed a pattern recognition technique to identify the more promising ARIMA models that should be considered for fitting to a specified time series. Following this, they recommended using an appropriate ASC to select the overall best model. Akaike (1977), Rissanen (1978) and Schwarz (1978) developed similar selection criteria for use with ARMA models while Chow (1978) proposed an improved version of these methods. Sawa (1978) defined a criterion for statistical model discrimination called the minimum attainable Bayes risk. Stone (1979) compared the asymptotic properties of the AIC and Schwarz criterion while Hannan (1979, 1980) derived important theoretical results for various kinds of ASC's. Based on the Kullback Leibler information number, Shibata (1989) obtained the TIC (Takeuchi's Information Criterion) as a natural extension of the AIC. He then went on to develop the RIC (Regularization Information Criterion) as a meaningful expansion of both the AIC and TIC. Moreover, Shibata (1989) compared various ASC's in terms of criteria which include consistency and efficiency.

As pointed out in Section 1.3.3, many of the ASC's have a structure which is quite similar to that of the AIC in [6.3.1]. Consider, for instance, Schwarz's approximation of the *Bayes information criterion (BIC)* (Schwarz, 1978) which is written as

$$BIC = -2\ln ML + k\ln(n) \qquad [6.3.5]$$

As is also the case for the AIC in [6.3.1], the first term on the right hand side of [6.3.5] reflects good statistical fit while the second component is concerned with model parsimony. When fitting more than one model to a given time series, one selects the model which gives the lowest value of the BIC. To employ an ASC such as the BIC in [6.3.5] in model construction, simply replace the AIC by the other ASC in Figure 6.3.1. As explained in Section 6.3.3, there are two basic approaches for utilizing an ASC in model development.

Certainly further theoretical and practical research is required to compare the capabilities of the more promising automatic selection procedures. However, the efficacy of MAICE is clearly demonstrated by the many and varied applications cited in this book and elsewhere. For instance, MAICE can be employed to choose the best model from different families of seasonal models (see Part VI), and to design transfer-function noise (Part VII), intervention (Part VIII) and multivariate ARMA (Part IX) models. Furthermore, when considering different types of models for forecasting, usually the kind of model which forecasts most accurately also possesses the lowest AIC value (see Chapter 15). Some disadvantages of MAICE and the other ASC's are that an overall statistic tends to cover up much of the information in the data and the practitioner may lose his or her sense of feeling for the inherent characteristics of the time series if he or she

bases his or her decisions solely upon one statistic. However, when MAICE is used in conjunction with the three stages of model construction as is shown in Figure 6.3.1, there is no doubt that MAICE greatly enhances the modelling process.

Akaike (1985) clearly explains how the derivation of the AIC is based upon the concept of *entropy*. In fact, the minimum AIC procedure can be considered as a realization of the *entropy-maximization principle* (Akaike, 1977). A further attractive theoretical feature of the MAICE approach is that it can be used to compare models which are not nested. Therefore, as noted earlier, one can use the MAICE procedure to select the best overall model across different families of models, as is done in Part VI for seasonal models. The practical import of the MAICE method for use in model discrimination is demonstrated by the two applications in the next section.

6.4 APPLICATIONS

6.4.1 Introduction

Table 5.4.1 in the previous chapter lists ARMA models identified for fitting to five nonseasonal stationary natural time series. In addition, Table 5.4.2 and Section 4.3.3 presents ARIMA models selected for fitting to three nonseasonal nonstationary time series. The maximum likelihood estimator described in Appendix A6.1 and mentioned in Section 6.2.3 can be used to calculate MLE's and SE's for the parameters in all of the foregoing models. Moreover, when more than one model is fitted to a given time series, the AIC of Section 6.3 can be employed for choosing the most appropriate model.

In the next two sections, estimation results along with applications of the AIC are presented for the same two case studies for which detailed identification findings are given in Section 5.4. The first application deals with modelling the average annual flows of the St. Lawrence River at Ogdensburg, New York, while the second one is concerned with modelling average annual sunspot numbers.

6.4.2 Yearly St. Lawrence Riverflows

Average annual flows for the St. Lawrence River at Ogdensburg, New York, are available from 1860 to 1957 (Yevjevich, 1963) and plotted in Figures 2.3.1 and 5.4.1 in m^3/s. The sample ACF, PACF, IACF, IPACF for these flows are displayed in Figures 5.4.2 to 5.4.5, respectively. As explained in Section 5.4.2, these identification graphs indicate that probably the best type of ARMA model to fit to the St. Lawrence flows is a constrained AR(3) model without the ϕ_2 parameter. However, one may also wish to try fitting AR(1) and unconstrained AR(3) models.

Table 6.4.1 lists the MLE's and SE's for AR(1), AR(3) and constrained AR(3) models fitted to the St. Lawrence flows. The theoretical definition for AR models can be found by referring to [3.2.5].

Model discrimination can be accomplished by comparing parameter estimates to their SE's, by using the AIC or by performing the likelihood ratio test. In order to employ the first procedure, first consider the models listed in Table 6.4.1. Notice that for both the AR(3) model and the AR(3) model without ϕ_2 the estimate $\hat{\phi}_3$ for ϕ_3 is more than twice its standard error. Therefore, it can be argued that even at the 1% significance level, ϕ_3 is significantly different from

Table 6.4.1. Parameter estimates for the AR models fitted to the annual St. Lawrence riverflows.

Models	Parameters	MLE's	SE's	AIC's
AR(1)	ϕ_1	0.708	0.072	1176.38
	σ_a	419.73		
AR(3)	ϕ_1	0.659	0.099	1175.59
	ϕ_2	-0.087	0.119	
	ϕ_3	0.216	0.099	
	σ_a	409.15		
Constrained AR(3) without ϕ_2	ϕ_1	0.619	0.084	1174.11
	ϕ_3	0.177	0.084	
	σ_a	410.27		

zero and should be included in the model. Consequently, the AR(1) model should not be utilized to model the St. Lawrence riverflows. Furthermore, because the SE for $\hat{\phi}_2$ in the AR(3) model is greater than $\hat{\phi}_2$, for model parsimony the AR(3) model without ϕ_2 is the proper model to select.

When the AIC is employed for model selection, it is not necessary to choose subjectively a significance level, as is done in hypothesis testing. By using [6.3.1], the values for the AIC are calculated for the three AR models and listed in the right hand column of Table 6.4.1. As can be seen, the AR(3) model without ϕ_2 has the minimum AIC, and, therefore, the AIC also indicates that this model should be chosen in preference to the others.

Suppose that one wishes to discriminate between models where one model is a subset of another. For the case of an AR model, let the order of one AR model be k and the order of another model containing more AR parameters be r. Let the residual variances of these two models be $\hat{\sigma}_a^2(k)$ and $\hat{\sigma}_a^2(r)$, respectively. The *likelihood ratio statistic* given by

$$n \ln \left[\hat{\sigma}_a^2(k) / \hat{\sigma}_a^2(r) \right] \sim \chi^2(r-k) \qquad [6.4.1]$$

is χ^2 distributed with $r-k$ degrees of freedom. If the calculated $\chi^2(r-k)$ from [6.4.1] is greater than $\chi^2(r-k)$ from the tables at a chosen significance level, a model with more parameters is needed.

The above likelihood ratio can be utilized to choose between the AR(1) model and the AR(3) model with $\phi_2 = 0$. By substituting $n = 97$, $k = 1$, the residual variance of the AR(1) model for $\hat{\sigma}_a^2(k)$, $r = 2$, and the residual variance of the AR(3) model with $\phi_2 = 0$ for $\hat{\sigma}_a^2(r)$, the calculated χ^2 statistic from [6.4.1] has a magnitude of 4.58. For 1 degree of freedom, this value is significant at the 5% significance level. Therefore, this test indicates that the constrained AR(3) model should be selected in preference to the AR(1) model.

The likelihood ratio test can also be employed to test whether an AR(3) model without ϕ_2 gives as good a fit as the AR(3) model. Simply substitute into [6.4.1] $n = 97$, $k = 2$, the residual variance of the AR(3) model with $\phi_2 = 0$ for $\hat{\sigma}_a^2(k)$, r =3, and the residual variance of the AR(3)

model for $\hat{\sigma}_a^2(r)$. The calculated χ^2 statistic possesses a value of 0.0569. For 1 degree of freedom this value is certainly not significant even at the 50% significance level. Consequently, the constrained model without ϕ_2 gives an adequate fit and should be used in preference to the AR(3) model in order to achieve model parsimony.

By substituting the estimated AR parameters into [3.2.5], one can write the constrained AR(3) model without ϕ_2 as

$$(1 - 0.619B - 0.177B^3)(z_t - 6819) = a_t \qquad [6.4.2]$$

where z_t is the average annual flow at time t, and 6819 is the MLE of the mean for the z_t series. Diagnostic checks presented in the next chapter in Section 7.6.2 demonstrate that the constrained AR(3) model without ϕ_2 adequately models the average annual flows of the St. Lawrence River.

6.4.3 Annual Sunspot Numbers

The yearly Wolfer sunspot number series is available from 1700 to 1960 (Waldmeier, 1961) where a plot of the series from 1770 to 1869 is shown in Figure 5.4.6. The sample ACF, PACF, IACF and IPACF are presented in Figures 5.4.7 to 5.4.10 in the identification chapter. As explained in Section 5.4.3, these identification graphs in conjunction with the output from diagnostic checks (see Section 7.6.3) indicate that an appropriate model may be a constrained AR(9) model without ϕ_3 to ϕ_8 fitted to the square roots of the sunspot series.

The MAICE procedure of Section 6.3 can be used to select the best type of ARMA model to fit to the sunspot series. Previously, Ozaki (1977) found using MAICE that an ARMA(6,3) model is the most appropriate model to fit to the given sunspot series having no data transformation. Akaike (1978) employed the AIC to select an ARMA(7,3) model with a square root transformation as the best sunspot model. However, Akaike (1978) did note that, because of the nature of sunspot activity, a model based on some physical consideration of the generating mechanism may produce a better fit to the data. Nevertheless, in this section it is shown how the model building procedure outlined in Figure 6.3.1 can be used to select an even better model from the family of ARMA models. As was suggested by McLeod et al. (1977) and also Hipel (1981), the AR(9) model, with a square-root transformation and the third to eighth AR parameters omitted from the model, produces a lower value of the AIC than all of the other aforementioned ARMA models. Earlier, Schaerf (1964) suggested modelling the sunspot data using a constrained AR(9) model but without the square-root transformation.

Because Ozaki (1977) used the series of 100 sunspot values listed as series E in the book of Box and Jenkins (1976), the same data set is used here for comparison purposes. By using an exhaustive enumeration procedure, Ozaki (1977) calculated the AIC for all ARMA(p,q) models for $0 \leq p,q \leq 9$ and found that an ARMA(6,3) model possessed the minimum AIC value. Employing [6.3.1], the values of the AIC were calculated for the same set of models examined by Ozaki (1977). The second column of Table 6.4.2 lists the AIC values for some of the models when the data is not transformed using [3.4.30] (i.e., λ=1 and c=0 in [3.4.30]). It can be seen that the minimum AIC value occurs for the ARMA(6,2) model, which is almost the same as the value of the AIC for the constrained AR(9) model. Notice that the ARMA(6,3) model suggested by Ozaki (1977) has a much higher AIC value than those for the ARMA(6,2) and constrained AR(9) models. This discrepancy is probably due to the different estimation procedure used by Ozaki.

The estimation method of McLeod (1977) described in Appendix A6.1 provides parameter estimates that are closer approximations than those of Box and Jenkins (1976) to the exact MLE's. As shown by McLeod (1977), implementation of his estimation method can result in improved estimates of the model parameters especially when MA parameters are contained in the model. As far as Table 6.4.2 is concerned, the improved estimation procedure affects the log likelihood in [6.3.1] and this in turn causes the AIC values to be slightly different than those given by Ozaki (1977, p. 297, Table 6).

Because information from the three stages of model construction is essentially ignored when using an exhaustive AIC enumeration such as the one adopted by Ozaki (1977), the best ARMA model is missed. To avoid this type of problem, the AIC can be combined with model construction as shown in Figure 6.3.1. From the plots of the sample ACF, PACF, IACF, and IPACF in Figures 5.4.7 to 5.4.10, respectively, it is difficult to decide upon which model to estimate. However, the sample PACF does possess values at lags 1 and 2 which are significantly different from zero and also some rather larger values at lags 6 to 9. When an ARMA(2,0) model is fitted to the data the independence, normality and homoscedastic assumptions (see Sections 7.3 to 7.5, respectively) are not satisfied. The residual ACF (see Section 7.3.2) has a large value at lag 9 and this fact suggests that an AR parameter at lag 9 should perhaps be incorporated into the model. The value of the AIC is lowest in column 2 of Table 6.4.2 for the AR(9) model without AR parameters from lags 3 to 8. However, because the statistic for changes in residual variance depending on the current level of the series, the statistic for trends in variance over time (see Section 7.5.2) and the skewness coefficient (see Section 7.4.2) all possess magnitudes which are more than twice their standard error, this points out the need for a Box-Cox transformation to eliminate heteroscedasticity and nonnormality. A square-root transformation can be invoked by setting λ equal to 0.5 in [3.4.30] and assigning the constant c a value of 1.0 due to the zero values in the sunspot series. Notice from the entries in the third column in Table 6.4.2 that a square-root transformation drastically lowers the AIC values for all of the models. The best model is an AR(9) or ARMA(9,0) model with a square-root transformation and without the third to eighth AR parameters. This constrained model was not missed because information from the model construction stages was used in conjunction with the MAICE procedure. Hence, when modelling a complex time series such as the sunspot data, it is advantageous for the practitioner to interact at all stages of model development by following the logic in Figure 6.3.1.

In Table 6.4.3, the MLE's and SE's are shown for the parameters of the most appropriate ARMA model which is fitted to the sunspot time series. When considering the 100 observations from 1770 to 1869 which are listed as Series E in Box and Jenkins (1976), the difference equation for the constrained AR(9) model with a square-root transformation is written as

$$(1 - 1.325B + 0.605B^2 - 0.130B^9)(w_t - 10.718) = a_t \qquad [6.4.3]$$

where

$$w_t = (1/0.5)[(z_t + 1.0)^{0.5} - 1.0]$$

is the transformation of the given z_t series for the sunspot numbers. The calibrated difference equation for the model fitted to the entire sunspot series from 1700 to 1960 is

Table 6.4.2. AIC values for the ARMA sunspot models.

ARMA (p,q) Model	AIC for λ=1, c=0.0	AIC for λ=0.5, c=1.0
(1,0)	618.30	580.40
(2,0)	551.85	518.41
(2,1)	547.63	519.19
(3,0)	549.57	519.13
(4,4)	546.98	516.10
(5,1)	547.29	523.82
(5,4)	548.20	517.96
(5,5)	547.23	517.38
(6,1)	548.11	517.39
(6,2)	545.05	518.43
(6,3)	551.35	523.34
(6,4)	550.73	502.40
(7,1)	548.17	518.98
(7,3)	551.01	521.37
(8,0)	545.67	519.83
(8,1)	547.65	520.44
(9,0)	547.65	519.72
(9,1)	548.18	518.78
Constrained (9,0)	545.64	511.58

$$(1 - 1.245B + 0.524B^2 - 0.192B^9)(w_t - 10.673) = a_t \qquad [6.4.4]$$

As shown in Section 7.6.3, the sunspot model in [6.4.4] satisfies diagnostic checks.

Table 6.4.3. Parameter estimates for the constrained AR(9) model fitted to the square roots of the yearly sunspot observations.

Parameters	MLE's	SE's
ϕ_1	1.325	0.074
ϕ_2	-0.605	0.076
ϕ_9	0.130	0.042
μ	10.718	1.417
σ_a^2	4.560	

By using [6.3.2] the relative plausibility of the sunspot models can be obtained. For instance, from Table 6.4.2 the next best model to the one in [6.4.3], according to the AIC, is an ARMA(4,4) model with a square-root transformation. When the appropriate AIC values from Table 6.4.2 are substituted into [6.3.2], the plausibility of the ARMA(4,4) model with a square-root transformation as compared to the best model is 0.10. According to the AIC, all of the other ARMA models with a square-root transformation are less plausible than even the ARMA(4,4) model. In addition, a comparison of the entries in columns two and three of Table 6.4.2 reveals how a square-root transformation significantly lowers the AIC values and hence increases the plausibility of a given ARMA model.

6.5 CONCLUSIONS

As explained in Section 6.2.2, maximum likelihood estimators possess a range of very desirable statistical properties which makes them highly attractive for use in practical applications. For example, maximum likelihood estimators are efficient and therefore produce parameter estimates having minimum variances in large samples. Accordingly, maximum likelihood estimation is the best approach for estimating the parameters in an ARMA model which is fitted to a given time series. Of particular, practical importance is the maximum likelihood estimator of McLeod (1977) described in Appendix A6.1 which is efficient both from statistical and computational viewpoints. This estimation procedure is used for estimating parameters in not only ARMA and ARIMA models but also many of the extensions to nonseasonal ARMA models presented later in the book and listed in Table 1.6.2.

Often the identification procedures of Section 5.3 suggest more than one model to fit to a specific time series. After calibrating the parameters for the ARMA models, the best overall model can be selected using the MAICE procedure of Section 6.3. The ways in which the MAICE approach can be incorporated into the three stages of model construction given in Figure III.I, are shown in Figure 6.3.1.

After selecting the best overall fitted model using the AIC or another appropriate ASC, the chosen model should be subjected to rigorous diagnostic checking. Procedures for making sure that various modelling assumptions are satisfied are described in detail in the next chapter.

APPENDIX A6.1

ESTIMATOR FOR ARMA MODELS

The purpose of this appendix is to describe the *modified sum of squares algorithm* of McLeod (1977) for obtaining approximate MLE's for the parameters in an ARMA model. As pointed out in Section 6.2.3 this estimator is computationally efficient and produces parameter estimates which are usually identical to the exact MLE's.

Let w_t, $t = 1,2,\ldots,n$, be a stationary time series which is normally distributed. One wishes to use the maximum likelihood estimator to obtain estimates for the parameters in the ARMA model defined in [3.4.4] or [4.3.4]. The parameters to estimate are:

1. the mean μ for the series. If the series has been differenced at least once, one may wish to set μ=0 for the w_t series. Otherwise, μ can be estimated using

$$\hat{\mu} = \bar{w} = \sum_{t=1}^{n} \frac{w_t}{n}$$

and then fixed at $\bar{w}$ when estimating the other model parameters. Another approach is to include μ as an additional parameter to estimate along with those mentioned below. For time series of moderate length (i.e., $n \geq 30$), the estimate given by $\bar{w}$ will be very close to that obtained when μ is iteratively estimated along with the other model parameters.

2. the p AR parameters contained in the set

$$\phi = (\phi_1, \phi_2, \ldots, \phi_p).$$

3. the q MA parameters in the set

$$\theta = (\theta_1, \theta_2, \ldots, \theta_q).$$

4. the innovation series given by $a_1, a_2, \ldots, a_n$.

5. the variance, σ_a^2, of the innovations.

To write down the likelihood function for an ARMA(p,q) model, one must assume a distribution for the innovations and hence the w_t series. In particular, assume the innovations are NID(0, σ_a^2) and the w_t sequence is $N(\mu, \sigma_w^2)$.

Recall that for a single random variable, w, which is $N(\mu, \sigma^2)$, the pdf is written as

$$p(w) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{\frac{-(w-\mu)^2}{2\sigma^2}\right\}$$

Suppose that one has a time series, w_t, of n random variables given by $w_1, w_2, \ldots, w_n$, where the w's are jointly normally distributed as

$$N(\mu, \Gamma_n^{(p,q)})$$

where

$$\mu^T = (\mu, \mu, \ldots, \mu)$$

since $\mu_1 = \mu_2 = \cdots = \mu_n$, $\Gamma_n^{(p,q)}(\phi,\theta) = E(\mathbf{w}\mathbf{w}^T)$ is the variance-covariance matrix of the w_t's where

$$\mathbf{w}^T = (w_1 - \mu, w_2 - \mu, \ldots, w_n - \mu)$$

and the (i,j) element of $\Gamma_n^{(p,q)}$ is $\gamma_{|i-j|}$ for which the autocovariance is defined by $\gamma_k = E[w_k, w_{t-k}]$ as in [2.5.3]. The joint normal distribution of the w_t's is given by

$$p(\mathbf{w}|\phi,\theta,\sigma_a^2,\mu) = (2\pi)^{-n/2}|\Gamma_n^{(p,q)}(\phi,\theta)|^{-1/2}\exp\left\{\frac{-\mathbf{w}^T(\Gamma_n^{(p,q)}(\phi,\theta))^{-1}\mathbf{w}}{2}\right\} \quad [A6.1.1]$$

Let $\mathbf{M}_n^{(p,q)}(\phi,\theta) = \dfrac{\sigma_a^2}{\Gamma_n^{(p,q)}(\phi,\theta)}$ and, hence, $[\Gamma_n^{(p,q)}(\phi,\theta)]^{-1} = \dfrac{\mathbf{M}_n^{(p,q)}(\phi,\theta)}{\sigma_a^2}$. Then,

$$p(\mathbf{w}|\phi,\theta,\sigma_a^2,\mu) = (2\pi\sigma_a^2)^{-n/2}|\mathbf{M}_n^{(p,q)}(\phi,\theta)|^{1/2}\exp\left\{\frac{-\mathbf{w}^T\mathbf{M}_n^{(p,q)}(\phi,\theta)\mathbf{w}}{2\sigma_a^2}\right\} \quad [A6.1.2]$$

When the unconditional sum of squares function is given by $S(\phi,\theta) = \sum_{t=-\infty}^{n} \hat{a}_t^2$, Box and Jenkins (1976, Ch. 7, A7.4) show that the above can be used to evaluate $\mathbf{w}^T\mathbf{M}_n^{(p,q)}(\phi,\theta)\mathbf{w}_n$ conveniently. The $\hat{a}_t$'s can be calculated using the back-forecasting procedure of Box and Jenkins (Box and Jenkins, 1976, pp. 215-220) as

$$\hat{a}_t = E[a_t|\mathbf{w},\phi,\theta]$$

More specifically, let $[w_t]$ and $[a_t]$ denote conditional expectations given $w_{t-1},w_{t-2},\ldots,w_1$. Then,

$$\phi(B)[w_t] = \theta(B)[a_t] \quad [A6.1.3]$$

where $[a_t] = 0, t > n$. Similarly,

$$\phi(F)[w_t] = \theta(F)[e_t] \quad [A6.1.4]$$

where F is the forward differencing operator defined by $Fw_t = w_{t+1}$, and $e_t \sim NID(0,\sigma_a^2)$ with $[e_t] = 0, t < 1$. Then, the unconditional sum of squares function $S(\phi,\theta)$ is calculated as follows:

Step 0: Initialization. Set Q large enough so the model is well approximated by a MA(Q) process. Typically, $Q \simeq 100$ is sufficient.

Step 1: Calculate $[w_t]$ $(t = n+Q,\ldots,1)$ using [A6.1.4]. Begin this calculation by setting $[w_t] = 0, t \geq n-p$.

Step 2: Calculate $[e_t]$ $(t = n+Q,\ldots,1)$ using [A6.1.3]. Start this calculation by setting $[e_t] = 0$, $t = n-p$.

Step 3: Back forecast w_t $(t = 0,-1,\ldots,1-Q)$. This is done by using [A6.1.4] to calculate first $[w_o]$ then $[w_{-1}],\ldots,[w_{1-Q}]$.

Step 4: Calculate $[a_t]$ $(t = 1-Q,\ldots,n)$ using [A6.1.3].

Step 5: $S(\phi,\theta) = \sum_{t=1-Q}^{n} [a_t]^2$.

Consequently, the likelihood function is given by

$$L(\phi,\theta,\sigma_a^2|w)\alpha\sigma_a^{-n}|M_n^{(p,q)}(\phi,\theta)|^{1/2}\exp\left\{-\frac{S(\phi,\theta)}{2\sigma_a^2}\right\} \qquad [A6.1.5]$$

Because the term $|M_n^{(p,q)}(\phi,\theta)|$ is dominated by the expression $\exp\{-S(\phi,\theta)/2\sigma_a^2\}$ in [A6.1.5] for large n and $|M_n^{(p,q)}(\phi,\theta)|$ is difficult to calculate, Box and Jenkins (1976, p. 213) suggest that the determinant can be disregarded and approximate MLE's can be obtained for the model parameters. However, if the sample is small and/or MA parameters are included in the model, the resulting parameter estimates may differ appreciably from the exact MLE's (McLeod, 1977). To rectify these problems various authors have suggested different approaches for calculating $|M_n^{(p,q)}(\phi,\theta)|$. McLeod (1977) devised a procedure whereby $|M_n^{(p,q)}(\phi,\theta)|$ is replaced by its asymptotic limit given by

$$m_{p,q}(\phi,\theta) = \lim_{n\to\infty}|M_n^{(p,q)}(\phi,\theta)| \qquad [A6.1.6]$$

When there are no MA parameters in an ARMA(p,q) model and hence $q = 0$, it is known that for $n \geq p$ (Box and Jenkins, 1976, p. 275)

$$|M_n^{(p,0)}(\phi)| = |M_p^{(p,0)}(\phi)|$$

and the matrix $M_p^{(p,0)}(\phi)$ has the (i,j)th element (Pagano, 1973; McLeod, 1977)

$$\sum_{k=0}^{min(i,j)}(\phi_{i-k-1}\phi_{j-k-1}-\phi_{p+1+k-i}\phi_{p+1+k-j})$$

where $\phi_0 = -1$. To calculate $m_{p,0}(\phi)$, one can use

$$m_{p,0}(\phi,\theta) = |M_p^{(p,0)}(\phi)|. \qquad [A6.1.7]$$

As shown by McLeod (1977)

$$m_{p,q}(\phi,\theta) = \frac{m_{p,0}^2(\phi)m_{q,0}^2(\theta)}{m_{p+q,0}(\phi^*)} \qquad [A6.1.8]$$

where ϕ^*_i is the ith parameter in the operator of order $p+q$ that is defined by

$$\phi^*(B) = \phi(B)\theta(B)$$

and $\phi^* = (\phi^*_1,\phi^*_2,\ldots,\phi^*_{p+q})$. Consequently, to compute $m_{p,q}(\phi,\theta)$ in [A6.1.8], it is only necessary to calculate the determinants of the three positive definite matrices which are obtained from [A6.1.7].

For convenience, McLeod (1977) defines the modified sum of squares function given by

$$S_m(\phi,\theta) = S(\phi,\theta)\{m_{p,q}(\phi,\theta)\}^{-1/n} \qquad [A6.1.9]$$

and this is the function called the modified sum of squares (MSS) referred to in Section 6.2.3. To obtain MLE's for the model parameters, the modified sum of squares must be minimized by using a standard optimization algorithm such as the method of Powell (1964, 1965). When modelling seasonal time series, it is a straightforward procedure to appropriately alter [A6.1.9] for use with seasonal ARMA models (McLeod and Sales, 1983).

As noted in Section 6.1, often the value of the Box-Cox parameter λ in [3.4.30] is known in advance for a given type of time series. If λ is not known, this parameter can be iteratively estimated along with the other ARMA model parameters. However, one must take into account the Jacobian of the transformation to obtain the log likelihood function given by (McLeod, 1974; Hipel et al., 1977)

$$l(\lambda,\phi,\theta,\sigma_a^2) \approx -\frac{n}{2}\ln\frac{MSS}{n} + (\lambda - 1)\sum_{t=1}^{n}\ln(w_t + c) \qquad \text{[A6.1.10]}$$

where c is the constant in the Box-Cox transformation in [3.4.30] that causes all entries in the w_t series to be positive. When all the entries in the w_t series are greater than zero, one sets $c = 0$. When λ is fixed beforehand or estimated, one should minimize [A6.1.10] to obtain MLE's for the model parameters. If a computer package does not possess the capability of obtaining the MLE of λ, the log likelihood can be calculated for a range of fixed values of λ, and the λ which gives the largest value of the log likelihood can be chosen.

When using the estimator of this appendix to obtain MLE's for an ARMA model or other types of models given in this text, it is recommended that the w_t series be standardized before using the estimator. For example, each observation in the w_t series can be standardized by subtracting out the mean of the series and dividing this by the standard deviation of the series. If the series is not standardized, one may run into numerical problems when optimizing the likelihood function. This is especially true for the transfer function-noise and intervention models in Parts VII and VIII, respectively where the absolute magnitude of an estimated transfer function parameter may be much greater than the absolute magnitudes of the AR and MA parameters contained in the correlated noise terms.

APPENDIX A6.2

INFORMATION MATRIX

To obtain SE's for the MLE's of the AR and MA parameters in an ARMA model, one must calculate the variance-covariance matrix for the model parameters. The square roots of the diagonal entries in this matrix constitute the SE's for the corresponding parameter estimates.

Because the variance-covariance matrix is the inverse of the Fisher information matrix, first consider the definition for the *information matrix*. Let the sets of AR and MA parameters given in Section 6.2.1 as $\phi = (\phi_1,\phi_2,\ldots,\phi_p)$ and $\theta = (\theta_1,\theta_2,\ldots,\theta_q)$, respectively, be included in a single set as $\beta = (\phi,\theta)$. The variance of the innovations is denoted by σ_a^2. The likelihood function is written as $L(\beta|\mathbf{w})$, where $\mathbf{w} = (w_1,w_2,\ldots,w_n)$ is the set of observations. Let

$$I(\beta) = \left[\lim_{n\to\infty} E\left\{-\frac{\partial^2 \ln L(\beta|w)}{\partial\beta_i \partial\beta_j}\Big|_{\beta=\hat{\beta}}/n\right\}\right] \qquad [A6.2.1]$$

where the (i,j) element is defined inside the brackets on the right hand side, the dimension of the information matrix is $(p+q)$ by $(p+q)$, β_i and β_j are the ith and jth parameters, respectively, and $\hat{\beta} = (\hat{\phi},\hat{\theta})$ is the set of MLE's for the AR and MA parameters. Then, $I(\beta)$ is said to be the theoretical Fisher large sample information per observation on β. In practice $I(\beta)$ is estimated by $I(\hat{\beta})$.

The *variance-covariance matrix* for $V(\hat{\beta})$ for the set of MLE's $\hat{\beta}$ is given in large samples by the inverse of the information matrix. Hence,

$$V(\hat{\beta}) \approx nI(\hat{\beta})^{-1} \qquad [A6.2.2]$$

The square roots of the diagonal entries in the variance-covariance matrix in [A6.2.2] provide the estimates for the *standard errors (SE's)* of the corresponding parameters. The variance-covariance matrix is often referred to as simply the *covariance matrix*.

The second order partial derivatives with respect to the model parameters reflect the rate of change of slope of the log likelihood function. When this slope change is high, there is less spread around an optimum point in the log likelihood function. This in turn means that the inverse of the slope change is small which indicates a smaller SE when considering a diagonal entry in the variance-covariance matrix.

For an ARMA model, the variance-covariance matrix can be written in terms of the AR and MA parameters. In practice, the entries in the matrix can be calculated numerically.

Because it is known that MLE's are asymptotically normally distributed, one can test whether or not a given MLE is significantly different from zero. For example, if zero falls outside the interval given by the MLE ± 1.96 SE, one can state that the estimate under consideration is significantly different from zero at the 5% significance level. If this were not the case, one may wish to omit this parameter from the model fitted to the series. Constrained models are described in Section 3.4.4 while an example of a constrained model is the constrained AR(3) model fitted to the yearly St. Lawrence riverflow in Sections 5.4.2, 6.4.2 and 7.6.2.

From the definition [A6.2.1] it may be shown that

$$I(\beta) = \begin{pmatrix} \gamma_{vv}(i-j) & \gamma_{vu}(i-j) \\ \gamma_{uv}(i-j) & \gamma_{uu}(i-j) \end{pmatrix} \qquad [A6.2.3]$$

where the (i,j) element in each partitioned matrix is indicated and $\gamma_{vv}(p\times p)$, $\gamma_{uu}(q\times q)$, $\gamma_{vu}(p\times q)$, $\gamma_{uv}(q\times p)$ are the theoretical auto and cross covariances defined by

$$\phi(B)v_t = -a_t,$$

$$\theta(B)u_t = a_t,$$

$$\gamma_{vv}(k) = E(v_t v_{t+k}),$$

$$\gamma_{uu}(k) = E(u_t u_{t+k}),$$

$$\gamma_{vu}(k) = E(v_t u_{t+k}),$$

$$\gamma_{uv}(k) = \gamma_{vu}(-k). \qquad \text{[A6.2.4]}$$

The covariance functions in [A6.2.4] may be obtained from a generalization of the algorithm given in Appendix A3.2.

APPENDIX A6.3

FINAL PREDICTION ERROR

Suppose that it is required to determine the order of an AR model to fit to a stationary time series $w_1, w_2, \ldots, w_n$. Prior to the introduction of the AIC defined in [6.3.1], Akaike (1969, 1970) developed a statistic called the *final prediction error (FPE)* for selecting the order of the AR model. The FPE is an estimate of the one step ahead prediction error variance of the AR(p) model in [3.2.5] and is defined as

$$FPE = \hat{\sigma}_a^2(p)\left[1 + \frac{p+1}{n}\right]\left[1 - \frac{p+1}{n}\right]^{-1} \qquad \text{[A6.3.1]}$$

where $\hat{\sigma}_a^2(p) = \frac{1}{n-p}\sum_{t=p+1}^{n} \hat{a}_t^2$ is the unbiased estimate of the residual variance of the AR(p) model. According to Akaike (1969, 1970), the AR model with the minimum value of the FPE in [A6.3.1] should be selected for modeling the series.

Taking natural logarithms of [A6.3.1] produces the result

$$\ln FPE = \ln\hat{\sigma}_a^2(p) + \frac{2(p+1)}{n} + 0(n^{-2}) \qquad \text{[A6.3.2]}$$

It is known that (-2) times the log likelihood of a Gaussian AR(p) model is approximately given by $n\ln\hat{\sigma}_a^2(p)$+ constant. Hence, as noted by Ozaki (1977),

$$n\ln FPE = AIC + constant + 0(n^{-1}) \qquad \text{[A6.3.3]}$$

Consequently, the MAICE procedure for AR model fitting is asymptotically equivalent to choosing the minimum value of the FPE.

PROBLEMS

6.1 Chapter 6 concentrates on explaining how the method of maximum likelihood can be used for estimating the parameters of ARMA models. However, other parameter estimation approaches are also available. Make a list of the names of six other estimation techniques. Outline the main ideas behind any two of these six methods.

6.2 In Section 6.2.2, first order and second order efficiency are referred to. Using equations where necessary, discuss these two concepts in more depth than that given in Section 6.2.2.

6.3 A criterion for characterizing an estimator is sufficiency. Define what is meant by sufficiency. Are maximum likelihood estimators sufficient?

6.4 What is an approximate maximum likelihood estimator? Outline the main components contained in the conditional and unconditional approximate maximum likelihood estimators suggested by Box and Jenkins (1976).

6.5 What is an exact maximum likelihoood estimator? Describe the main steps followed when applying the exact maximum likelihood estimators provided by Ansley (1979) as well as Ljung and Box (1979).

6.6 To optimize a likelihood or log likelihood function, a number of optimization algorithms are listed in Section 6.2.3. Outline the steps contained in the conjugate directions algorithm of Powell (1964, 1965). Discuss the advantages and limitations of Powell's algorithm.

6.7 Explain the difference between maximum likelihood estimation and Gaussian estimation.

6.8 Show that the exact log likelihood function for a Gaussian AR(1) is:

$$z_t = \mu + \phi_1(z_{t-1} - \mu) + a_t$$

where $a_t \sim NID(0,\sigma_a^2)$ and $t = 1, \ldots, n$ may be written as

$$\log L(\phi_1,\mu,\sigma_a^2) = -\frac{n}{2}\log\sigma_a^2 + \frac{1}{2}\log(1-\phi_1^2) - \frac{1}{2\sigma_a^2}S(\phi_1)$$

where

$$S(\phi_1) = (1-\phi_1^2)(z_1-\mu)^2 + \sum_{t=2}^{n}[(z_t-\mu) - \phi_1(z_{t-k}-\mu)]^2$$

Simulate z_t, $t = 1, \ldots, n$ several times for various n and plot $L(\phi_1,\mu,\sigma_a^2)$ for $|\phi_1| < 1$.

6.9 Suppose that

$$(1-\phi_1 B)z_t = a_t$$

where $\log a_t \sim NID(0,\sigma_a^2)$. Show that $I(\phi_1) = (1+\sigma_a^{-2})e^{2\sigma_a^2}$

(a) Consider the AR(1) model

$$(1-\phi B)z_t = a_t$$

If $a_t \sim NID(0,1)$, show that $I(\phi) = \dfrac{1}{1-\phi^2}$

(b) Now consider the AR(1) model

$$(1-\phi B)z_t = a_t,$$

where $\log a_t \sim NID(0,1)$. Show that in this case, that

$$I(\phi) = 2e^2\left[\frac{e(e-1)}{1-\phi^2} + \frac{e}{(1-\phi)^2}\right]$$

(c) Compare the relative efficiency of Gaussian estimation versus maximum likelihood estimation when $\log a_t \sim NID(0,1)$. Verify your theoretical calculation by simulation (see Chapter 9 for an explanation of simulation).

6.10 Outline the theoretical development of the AIC given in [6.3.1].

6.11 Two approaches for employing the AIC in conjunction with model construction are described in Section 6.3.3. Using an annual time series of your choice, employ these two procedures for determining the best overall ARMA or ARIMA model to fit to the series.

6.12 Compare the MA(2) and AR(2) models for the Mean Annual Temperatures in Central England. Calculate the plausibility of the MA(2) model versus the AR(2) model.

6.13 The general form of an automatic selection criterion for model discrimination is given in Section 1.3.3 while the AIC and BIC are defined in [6.3.1] and [6.3.5]. Excluding the AIC and BIC, give the definitions of three other ASC's. Discuss the domains of applicability, advantages and drawbacks of each of these ASC's.

REFERENCES

AKAIKE INFORMATION CRITERION (AIC)

Akaike, H. (1971). Determination of the number of factors by an extended maximum likelihood principle. Research memorandum No. 44, The Institute of Statistical Mathematics, Tokyo.

Akaike, H. (1972a). Use of an information theoretic quantity for statistical model identification. In *Proceedings of the 5th Hawaii International Conference on Systems Sciences*, 249-250, Western Periodicals, North Hollywood, California.

Akaike, H. (1972b). Automatic data structure by maximum likelihood. In *Computers in Biomedicine*, Supplement to *Proceedings of 5th International Conference on Systems Sciences*, 99-101, Western Periodicals, North Hollywood, California.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csadki, F., Editors, *Proceedings of the 2nd International Symposium on Information Theory*, 267-281, Budapest. Akademial Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716-723.

Akaike, H. (1976). Canonical correlation analysis of time series and the use of an information criterion. In Mehra, R. K. and Lainiotis, D. G., Editors, *System Identification*, 27-96. Academic Press, New York.

Akaike, H. (1978). On the likelihood of a time series model. Paper presented at the Institute of Statisticians 1978 Conference on Time Series Analysis and Forecasting, Cambridge University.

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika* 66(2):237-242.

Akaike, H. (1985). Prediction and entropy. In Atkinson, A. C. and Fineburg, F. E., Editors, *A Celebration of Statistics*, 1-24, Springer-Verlag, Berlin.

Hipel, K. W. (1981). Geophysical model discrimination using the Akaike information criterion. *IEEE Transactions on Automatic Control*, AC-26(2):358-378.

Hipel, K. W., McLeod, A. I. and Lennox, W. C. (1977a). Advances in Box-Jenkins modelling, 1, Model construction. *Water Resources Research*, 13(3):567-575.

Kitagawa, G. (1979). On the use of AIC for detection of outliers. *Technometrics*, 21(2):193-199.

McLeod, A. I. and Hipel, K. W. (1978). Preservation of the rescaled adjusted range, 1, A reassessment of the Hurst phenomenon. *Water Resources Research*, 14(3):491-508.

McLeod, A. I., Hipel, K. W. and Lennox, W. C. (1977). Advances in Box-Jenkins modelling, 2, Applications. *Water Resources Research*, 13(3):577-586.

Ozaki, T. (1977). On the order determination of ARIMA models. *Journal of the Royal Statistical Society*, Series C (Applied Statistics), 26(3):290-301.

Sakamoto, Y. and Akaike, H. (1977). Analysis of cross classified data by AIC. *Annals of the Institute of Statistical Mathematics*, B:30-31.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, 63:117-126.

Tanabe, K. (1974). Fitting regression curves and surfaces by Akaike's Information Criterion. Research Memo, No. 63, The Institute of Statistical Mathematics, Tokyo.

Tong, H. (1975). Determination of the order of a Markov chain by Akaike's information criterion. *Journal of Applied Probability*, 12(3):488-497.

AUTOMATIC SELECTION CRITERIA BESIDES AIC

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21:243-247.

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, 22:203-217.

Akaike, H. (1977). On entropy maximization principle. In Krishnaiah, P. R., Editor, *Applications of Statistics*, 27-41. North-Holland, Amsterdam.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. John Wiley, New York.

Chow, G. C. (1978). A reconcilliation of the information and posterior probability criteria for model selection. Research Memorandum No. 234, Econometric Research Program, Princeton University.

Gray, H. L., Kelley, G. D. and McIntire, D. D. (1978). A new approach to ARMA modelling. *Communications in Statistics*, B7(1):1-77.

Hannan, E. J. (1979). Estimating the dimension of a linear system. Unpublished manuscript, The Australian National University, Canberra, Australia.

Hannan, E. J. (1980). The estimation of the order of an ARMA process. *Annals of Statistics*, 8:1071-1081.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, Series B, 41(2):190-195.

Hill, G. W. and Woodworth, D. (1980). Automatic Box-Jenkins forecasting. *Journal of the Operational Research Society*, 31(5):413-422.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15:661-675.

McClave, J. T. (1975). Subset autoregression. *Technometrics*, 17(2):213-220.

McClave, J. T. (1978). Estimating the order of autoregressive models: The Max χ^2 method. *Journal of the American Statistical Association*, 73(363):122-128.

Parzen, E. (1974). Some recent advances in time series modelling. *IEEE Transactions on Automatic Control*, AC-19(6):723-730.

Rissanen, J. (1978). Modelling by shortest data description. *Automatica*, 14:465-471.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica*, 46(6):1273-1291.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461-464.

Shibata, R. (1989). Statistical aspects of model selection. Working Paper WP-89-077, International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria.

Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society*, Series B, 41(2):276-278.

DATA SETS

Waldmeier, M. (1961). *The Sunspot Activity in the Years 1610-1960*. Schulthas and Company, Zurich, Switzerland.

Yevjevich, V. M. (1963). Fluctuation of wet and dry years, 1, Research data assembly and mathematical models. Hydrology Paper No. 1, Colorado State University, Fort Collins, Colorado.

ESTIMATION

Ansley, C. F. (1979). An algorithm for the exact likelihood of a mixed autoregressive-moving average process. *Biometrika*, 66:59-65.

Barnard, G. A. (1949). Statistical inference. *Journal of the Royal Statistical Society*, Series B, 11:115-149.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57:269-326.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, Oakland, California, revised edition.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Cramer, H. (1946). *Mathematical Models of Statistics*. Princeton University Press.

Edwards, A. W. F. (1972). *Likelihood*. Cambridge University Press, Cambridge, United Kingdom.

Fisher, R. A. (1922). On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society*, Series A, 222:308-358.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22:700-725.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

Hannan, E. J. (1970). *Multiple Time Series*. John Wiley, New York.

Kalman, R. E. (1960). A new approach of linear filtering and prediction problems. *Journal of Basic Engineering, Transactions ASME*, Series D, 82:35-45.

Kempthorne, O. and Folks, L. (1971). *Probability, Statistics and Data Analysis*. The Iowa State University Press, Ames, Iowa.

Kendall, M. G. and Buckland, W. R. (1971). *A Dictionary of Statistical Terms*. Longman Group Limited, Thetford, Norfolk, Great Britain, third edition.

Kotz, S. and Johnson, N. L., Editors (1988). *Encyclopedia of Statistical Sciences, Volumes 1 to 9*. Wiley, New York.

Kruskal, W. H. and Tanur, J. M. (1978). *International Encyclopedia of Statistics, Volumes 1 and 2*. The Free Press, New York.

Li, W. K. and McLeod, A. I. (1988). ARMA modelling with non-Gaussian innovations. *Journal of Time Series Analysis*, 9(2):155-168.

Ljung, G. M. and Box, G. E. P. (1979). The likelihood function of stationary autoregressive-moving average models. *Biometrika*, 66(2):265-270.

McLeod, A. I. (1977). Improved Box-Jenkins estimators. *Biometrika*, 64(3):531-534.

McLeod, A. I. and Sales, P. R. H. (1983). An algorithm for approximate likelihood calculation of ARMA and seasonal ARMA models. *Journal of the Royal Statistical Society, Series C* (Applied Statistics), 32:211-223.

Mélard, G. (1984). Algorithm AS197. A fast algorithm for the exact likelihood of autoregressive-moving average models. *Journal of the Royal Statistical Society, Part C, Applied Statistics*, 33:104-114.

Mendel, J. M. (1987). *Lessons in Digital Estimation Theory*. Prentice-Hall, Englewood Cliffs, New Jersey.

Newbold, P. (1974). The exact likelihood function for a mixed autoregressive-moving average process. *Biometrika*, 61(3):423-426.

Norden, R. H. (1972). A survey of maximum likelihood estimation. *International Statistical Review*, 40(3):329-354.

Norden, R. H. (1973). A survey of maximum likelihood estimation, part 2. *International Statistical Review*, 41(1):39-58.

Pagano, M. (1973). When is an autoregressive scheme stationary. *Communications in Statistics*, 1(6):533-544.

Rao, C. R. (1961). Asymptotic efficiency and limiting information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:531-546, University of California, Berkeley.

Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples (with discussion). *Journal of the Royal Statistical Society, Series B*, 24:46-72.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley, New York, second edition.

Sachs, L. (1984). *Applied Statistics, A Handbook of Techniques*. Springer-Verlag, New York, second edition.

Schaerf, M. C. (1964). Estimation of the covariance autoregressive structure of a stationary time series. Technical report, Department of Statistics, Stanford University, Stanford, California.

Whittle, P. (1961). Gaussian estimation in stationary time series. *Bulletin of the International Statistical Institute*, 39:105-128.

OPTIMIZATION

Davidon, W. C. (1968). Variance algorithm for minization. *The Computer Journal*, 10:406-410.

Draper, N. R. and Smith, H. (1980). *Applied Regression Analysis*. Wiley, New York, second edition.

Gill, P., Murray, W. and Wright, M. (1981). *Practical Optimization*. Academic Press, New York.

Ishiguro, M. and Akaike, H. (1989). DALL: Davidon's algorithm for log likelihood maximization - a FORTRAN subroutine for statistical model builders. *The Institute of Statistical Mathematics*, Tokyo, Japan.

Luenberger, D. G. (1984). *Linear and Nonlinear Programming*. Addison-Wesley, Reading, Massachusetts, second edition.

Marquardt, D. W. (1963). An algorithm for least squares estimation of nonlinear parameters. *Journal of the Society of Industrial and Applied Mathematics*, 11(2):431-441.

Powell, M. J. D. (1964). An efficient method for finding the minimum of a function of several variables with calculating derivatives. *Computer Journal*, 7:155-162.

Powell, M. J. D. (1965). A method for minimizing a sum of squares of nonlinear functions without calculating derivatives. *Computer Journal*, 8:303-307.

Vanderplaats, G. N. (1984). *Numerical Optimization Techniques for Engineering Design with Applications*. McGraw-Hill, New York.