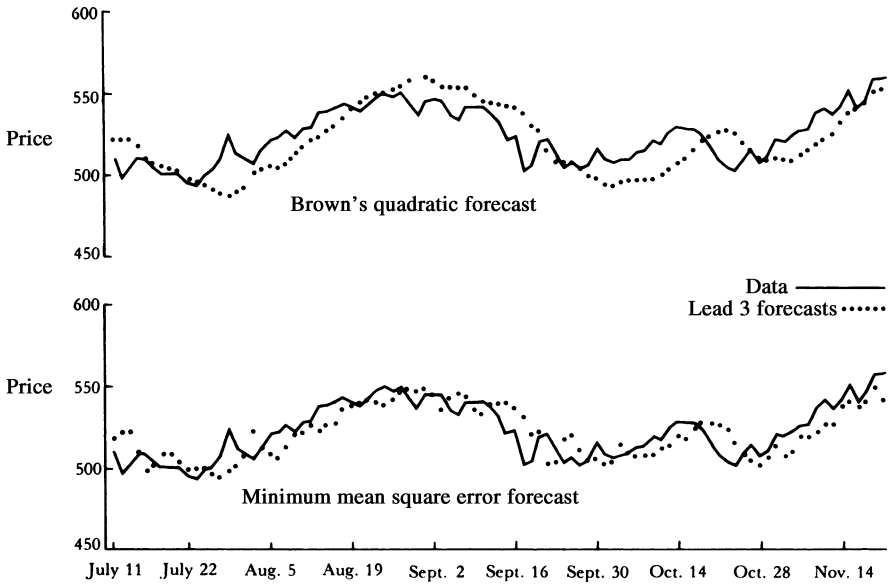

FIGURE A5.2 IBM stock price series with comparison of lead 3 forecasts obtained from best IMA(0, 1, 1) model and Brown's quadratic forecast for a period beginning from July 11, 1960.

which, arguing as in Section 4.3.3, can be written as

$$\nabla^3 z_t = \nabla^3 a_t + \lambda_0 \nabla^2 a_{t-1} + \lambda_1 \nabla a_{t-1} + \lambda_2 a_{t-1}$$

However, we show in Chapter 7 that if this model is correctly fitted, the least-squares estimates of the parameters are $\lambda_1 = \lambda_2 = 0$ and $\lambda_0 \simeq 1.0$. Thus, $\nabla z_t = (1 - \theta B)a_t$, with $\theta = 1 - \lambda_0$ close to zero, is the appropriate stochastic model, and the appropriate forecasting polynomial is $\hat{z}_t(l) = \beta_0^{(t)}$, which is of degree 0 in l and not of degree 2.

2. The choice of the weight function ω_j in (A5.3.8) must correspondingly be decided by the stochastic model, and not arbitrarily. The use of the discounted least-squares fitting procedure would produce minimum mean square error forecasts in the very restricted case, where

a. the process was of order (0, 1, 1), so $\nabla z_t = (1 - \theta B)a_t$,

b. a polynomial of degree 0 was fitted, and

c. the smoothing constant α was set equal to our $\lambda = 1 - \theta$.

In the present example, even if the correct polynomial model of degree 0 had been chosen, the value $\alpha = \lambda = 0.1$, actually used by Brown, would have been quite inappropriate. The correct value λ for this series is close to unity.

3. The exponentially discounted weighted least-squares procedure forces all the $p + d$ coefficients in the updating vector $\mathbf{h}$ to be functions of the single smoothing parameter α. In fact, they should be functions of the $p + q$ independent parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$.

Thus, the differences between the two methods are not trivial, and it is interesting to compare their performances on the IBM data. The minimum mean square error forecast is

TABLE A5.3 Comparison of Mean Square Error of Forecasts Obtained at Various Lead Times Using Best IMA(0, 1, 1) Model and Brown's Quadratic Forecasts

	Lead Time l									
	1	2	3	4	5	6	7	8	9	10
MSE (Brown)	102	158	218	256	363	452	554	669	799	944
MSE ($\lambda = 0.9$)	42	91	136	180	282	266	317	371	427	483

$\hat{z}_t(l) = b_0(t)$, with updating $b_0^{(t)} = b_0^{(t-1)} + \lambda a_t$, where $\lambda \simeq 1.0$. If λ is taken to be exactly equal to unity, this is equivalent to using

$$\hat{z}_t(l) = z_t$$

which implies that the best forecast of the stock price for all future time is the present price.[1] The suggestion that stock prices behave in this way is, of course, not new and goes back to Bachelier (1900). Since $z_t = Sa_t$ when $\lambda = 1$, this implies that z_t is a random walk.

To compare the minimum mean square error forecast with Brown's quadratic forecasts, a direct comparison was made using the IBM stock price series from July 11, 1960 to February 10, 1961, for 150 observations. For this stretch of the series, the minimum MSE forecast is obtained using the model $\nabla z_t = a_t - \theta a_{t-1}$, with $\theta = 0.1$, or $\lambda = 1 - \theta = 0.9$. Figure A5.2 shows the minimum MSE forecasts for lead time 3 and the corresponding values of Brown's quadratic forecasts. It is seen that the minimum MSE forecasts, which are virtually equivalent to using today's price to predict that 3 days ahead, are considerably better than those obtained using Brown's more complicated procedure.

The mean square errors for the forecast at various lead times, computed by direct comparison of the value of the series and their lead l forecasts, are shown in Table A5.3 for the two types of forecasts. It is seen that Brown's quadratic forecasts have mean square errors that are much larger than those obtained by the minimum mean square error method.

EXERCISES

5.1. For the models

(1) $\tilde{z}_t - 0.5\tilde{z}_{t-1} = a_t$

(2) $\nabla z_t = a_t - 0.5a_{t-1}$

(3) $(1 - 0.6B)\nabla z_t = a_t$

write down the forecasts for lead times $l = 1$ and $l = 2$:

(a) From the difference equation

(b) In integrated form (using the ψ_j weights)

(c) As a weighted average of previous observations

[1] This result is approximately true supposing that no relevant information except past values of the series itself is available and that fairly short forecasting periods are being considered. For longer periods, growth and inflationary factors would become important.

5.2. The following observations represent values $z_{91}, z_{92}, \ldots, z_{100}$ from a series fitted by the model $\nabla z_t = a_t - 1.1a_{t-1} + 0.28a_{t-2}$:

$$166, 172, 172, 169, 164, 168, 171, 167, 168, 172$$

(a) Generate the forecasts $\hat{z}_{100}(l)$ for $l = 1, 2, \ldots, 12$ and draw a graph of the series values and the forecasts (assume $a_{90} = 0, a_{91} = 0$).

(b) With $\hat{\sigma}_a^2 = 1.103$, calculate the estimated standard deviations $\hat{\sigma}(l)$ of the forecast errors and use them to calculate 80% probability limits for the forecasts. Insert these probability limits on the graph, on either side of the forecasts.

5.3. Suppose that the data of Exercise 5.2 represent monthly sales.

(a) Calculate the minimum mean square error forecasts for quarterly sales for 1, 2, 3, 4 *quarters* ahead, using the data up to $t = 100$.

(b) Calculate 80% probability limits for these forecasts.

5.4. Using the data and forecasts of Exercise 5.2, and given the further observation $z_{101} = 174$:

(a) Calculate the forecasts $\hat{z}_{101}(l)$ for $l = 1, 2, \ldots, 11$ using the updating formula

$$\hat{z}_{t+1}(l) = \hat{z}_t(l+1) + \psi_l a_{t+1}$$

(b) Verify these forecasts using the difference equation directly.

5.5. For the model $\nabla z_t = a_t - 1.1a_{t-1} + 0.28a_{t-2}$ of Exercise 5.2:

(a) Write down expressions for the forecast errors $e_t(1), e_t(2), \ldots, e_t(6)$, from the same origin t.

(b) Calculate and plot the autocorrelations of the series of forecast errors $e_t(3)$.

(c) Calculate and plot the correlations between the forecast errors $e_t(2)$ and $e_t(j)$ for $j = 1, 2, \ldots, 6$.

5.6. Let the vector $\boldsymbol{e}' = (e_1, e_2, \ldots, e_L)$ have for its elements the forecast errors made $1, 2, \ldots, L$ steps ahead, all from the same origin t. Then if $\boldsymbol{a}' = (a_{t+1}, a_{t+2}, \ldots, a_{t+L})$ are the corresponding uncorrelated random shocks, show that

$$\boldsymbol{e} = \mathbf{M}\boldsymbol{a} \quad \text{where} \quad \mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \psi_1 & 1 & 0 & \cdots & 0 \\ \psi_2 & \psi_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \psi_{L-1} & \psi_{L-2} & \psi_{L-3} & \cdots & 1 \end{bmatrix}$$

Also, show that (e.g., Box and Tiao, 1976; Tiao et al., 1975) $\boldsymbol{\Sigma}_e$, the covariance matrix of the e's, is $\boldsymbol{\Sigma}_e = \sigma_a^2 \mathbf{M}\mathbf{M}'$ and hence that a test to determine if a set of subsequently realized values $z_{t+1}, z_{t+2}, \ldots, z_{t+L}$ of the series taken jointly differ significantly from the forecasts made at the origin t is obtained by referring

$$\boldsymbol{e}'\boldsymbol{\Sigma}_e^{-1}\boldsymbol{e} = \frac{\boldsymbol{e}'(\mathbf{M}\mathbf{M}')^{-1}\boldsymbol{e}}{\sigma_a^2} = \frac{\boldsymbol{a}'\boldsymbol{a}}{\sigma_a^2} = \frac{1}{\sigma_a^2}\sum_{j=1}^{L} a_{t+j}^2$$

to a chi-square distribution with L degrees of freedom. Note that a_{t+j} is the *one*-step-ahead forecast error calculated from $z_{t+j} - \hat{z}_{t+j-1}(1)$.

5.7. Suppose that a quarterly economic time series is well represented by the model

$$\nabla z_t = 0.5 + (1 - 1.0B + 0.5B^2)a_t$$

with $\sigma_a^2 = 0.04$.

(a) Given $z_{48} = 130, a_{47} = -0.3, a_{48} = 0.2$, calculate and plot the forecasts $\hat{z}_{48}(l)$ for $l = 1, 2, \ldots, 12$.

(b) Calculate and insert the 80% probability limits on the graph.

(c) Express the series and forecasts in integrated form.

5.8. Consider the annual Wölfer sunspot numbers for the period 1770–1869 listed as Series E in Part Five of this text. The same series is available for the longer period 1700–1988 as "sunspot.year" in the datasets package of R. You can use either data set. Suppose that the series can be represented by an autoregressive model of order 3.

(a) Plot the time series and comment. Does the series look stationary?

(b) Generate forecasts and associated probability limits for up to 20 time periods ahead for the series.

(c) Perform a square root transformation of the data and repeat (a) and (b) above.

(d) Use the function BoxCox.ar() in the TSA package of R to show that the square root transformation is appropriate for this series; see help(BoxCox.ar) for details. (*Note:* Adding a small amount, for example, 1/2, to the series, eliminates zero values and allows the program to consider a log transformation as an option).

5.9. A time series representing a global mean land–ocean temperature index from 1880 to 2009 is available in a file called ''gtemp'' in the astsa package of R. The data are temperature deviations, measured in degree centigrades, from the 1951–1980 average temperature, as described by Shumway and Stoffer (2011, p. 5). Assume that a third-order autoregressive model is appropriate for the first differences $w_t = (1 - B)z_t$ of this series.

(a) Plot the time series z_t and the differenced series w_t using R.

(b) Generate forecasts and associated probability limits for up to 20 time periods ahead for this series using the function sarima.for() without including a constant term in the model.

(c) Generate the same forecasts and probability limits as in part (b) but with a constant term now added to the model. Discuss your findings and comment on the implications of including a constant in this case.

5.10. For the model $(1 - 0.6B)(1 - B)z_t = (1 + 0.3B)a_t$, express explicitly in the state-space form of (5.5.2) and (5.5.3), and write out precisely the recursive relations of the Kalman filter for this model. Indicate how the (exact) forecasts $\hat{z}_{t+l|t}$ and their forecast error variances $v_{t+l|t}$ are determined from these recursions.

PART TWO

STOCHASTIC MODEL BUILDING

We have seen that an ARIMA model of order (p, d, q) provides a class of models capable of representing time series that, although not necessarily stationary, are homogeneous and in statistical equilibrium in many respects.

The ARIMA model is defined by the equation

$$\phi(B)(1 - B)^d z_t = \theta_0 + \theta(B)a_t$$

where $\phi(B)$ and $\theta(B)$ are operators in B of degree p and q, respectively, whose zeros lie outside the unit circle. We have noted that the model is very general, including as special cases autoregressive models, moving average models, mixed autoregressive–moving average models, and the integrated forms of all three.

Iterative Approach to Model Building. The development of a model of this kind to describe the dependence structure in an observed time series is usually best achieved by a three-stage iterative procedure based on identification, estimation, and diagnostic checking.

1. By *identification* we mean the use of the data, and of any information on how the series was generated, to suggest a subclass of parsimonious models worthy to be entertained.
2. By *estimation* we mean efficient use of the data to make inferences about the parameters conditional on the adequacy of the model entertained.
3. By *diagnostic checking* we mean checking the fitted model in its relation to the data with intent to reveal model inadequacies and so to achieve model improvement.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

In Chapter 6, which follows, we discuss model identification, in Chapter 7 estimation of parameters, and in Chapter 8 diagnostic checking of the fitted model. In Chapter 9 we expand on the class of models developed in Chapters 3 and 4 to the *seasonal* ARIMA models, and all the model building techniques of the previous chapters are illustrated by applying them to modeling and forecasting seasonal time series. In Chapter 10 we consider some additional topics that represent extensions beyond the linear ARIMA class of models such as conditional heteroscedastic time series models, nonlinear time series models, and fractionally integrated long memory processes, which allow for certain more general features in the time series than are possible in the linear ARIMA models. Unit root testing is also discussed in this chapter.

6

MODEL IDENTIFICATION

In this chapter, we discuss methods for identifying nonseasonal autoregressive integrated moving average (ARIMA) time series models. Identification methods are rough procedures applied to a set of data to indicate the kind of model that is worthy of further investigation. The specific aim here is to obtain some idea of the values of p, d, and q needed in the general linear ARIMA model and to obtain initial estimates for the parameters. The tentative model specified provides a starting point for the application of the more formal and efficient estimation methods described in Chapter 7. The examples used to demonstrate the model-building process will include Series A–F that have been discussed in earlier chapters and are listed in the Collection of Time Series in Part Five of this book. The series are also available electronically at http://pages.stat.wisc.edu/ reinsel/bjr-data/.

6.1 OBJECTIVES OF IDENTIFICATION

It should first be said that identification and estimation necessarily overlap. Thus, we may estimate the parameters of a model, which is more elaborate than the one we expect to use, so as to decide *at what point* simplification is possible. Here we employ the estimation procedure to carry out part of the identification. It should also be explained that identification is necessarily inexact. It is inexact because the question of what types of models occur in practice and in what specific cases depends on the behavior of the physical world and therefore cannot be decided by purely mathematical argument. Furthermore, because at the identification stage no precise formulation of the problem is available, statistically ''inefficient'' methods must necessarily be used. It is a stage at which graphical methods are particularly useful and judgment must be exercised. However, it should be kept in mind

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

that the preliminary identification commits us to nothing except tentative consideration of a class of models that will later be efficiently fitted and checked.

6.1.1 Stages in the Identification Procedure

Our task, then, is to identify an appropriate subclass of models from the general ARIMA family

$$\phi(B)\nabla^d z_t = \theta_0 + \theta(B)a_t \tag{6.1.1}$$

which may be used to represent a given time series. Our approach will be as follows:

1. To assess the stationarity of the process z_t and, if necessary, to difference z_t as many times as is needed to produce stationarity, hopefully reducing the process under study to the mixed autoregressive–moving average process:

   $$\phi(B)w_t = \theta_0 + \theta(B)a_t$$

   where

   $$w_t = (1 - B)^d z_t = \nabla^d z_t$$

2. To identify the resulting autoregressive–moving average (ARMA) model for w_t.

Our principal tools for putting steps 1 and 2 into effect will be the sample autocorrelation function and the sample partial autocorrelation function. They are used not only to help guess the form of the model but also to obtain approximate estimates of the parameters. Such approximations are often useful at the estimation stage to provide starting values for iterative procedures employed at that stage. Some additional model identification tools may also be employed and are discussed in Section 6.2.4.

6.2 IDENTIFICATION TECHNIQUES

6.2.1 Use of the Autocorrelation and Partial Autocorrelation Functions in Identification

Identifying the Degree of Differencing. We have seen in Section 3.4.2 that for a stationary mixed autoregressive–moving average process of order $(p, 0, q)$, $\phi(B)\tilde{z}_t = \theta(B)a_t$, the autocorrelation function satisfies the difference equation

$$\phi(B)\rho_k = 0, k > q$$

Also, if $\phi(B) = \prod_{i=1}^{p}(1 - G_i B)$, the solution of this difference equation for the kth autocorrelation is, assuming distinct roots, of the form

$$\rho_k = A_1 G_1^k + A_2 G_2^k + \cdots + A_p G_p^k \qquad k > q - p \tag{6.2.1}$$

The stationarity requirement that the zeros of $\phi(B)$ lie outside the unit circle implies that the roots $G_1, G_2, \ldots, G_p$ lie inside the unit circle.

This expression shows that in the case of a stationary model in which none of the roots lie close to the boundary of the unit circle, the autocorrelation function will quickly ''die out'' for moderate and large k. Suppose now that a single real root, say G_1, approaches unity, so that

$$G_1 = 1 - \delta$$

where δ is some small positive quantity. Then, since for k large

$$\rho_k \simeq A_1(1 - k\delta)$$

the autocorrelation function will not die out quickly and will fall off slowly and very nearly linearly. A similar argument may be applied if more than one of the roots approaches unity.

Therefore, a tendency for the autocorrelation function not to die out quickly is taken as an indication that a root close to unity may exist. The estimated autocorrelation function tends to follow the behavior of the theoretical autocorrelation function. Therefore, failure of the estimated autocorrelation function to die out rapidly might logically suggest that we should treat the underlying stochastic process as nonstationary in z_t, but possibly as stationary in ∇z_t, or in some higher difference.

However, even though failure of the estimated autocorrelation function to die out rapidly suggests nonstationarity, the estimated autocorrelations need not be extremely high even at low lags. This is illustrated in Appendix A6.1, where the expected behavior of the estimated autocorrelation function is considered for the nonstationary $(0, 1, 1)$ process $\nabla z_t = (1 - \theta B)a_t$. The ratio $E[c_k]/E[c_0]$ of expected values falls off only slowly, but depends initially on the value of θ and on the number of observations in the series, and need not be close to unity if θ is close to 1. We illustrate this point again in Section 6.3.4 for Series A.

For the reasons given, it is assumed that the degree of differencing d, necessary to achieve stationarity, has been reached when the autocorrelation function of $w_t = \nabla^d z_t$ dies out fairly quickly. In practice, d is normally 0, 1, or 2, and it is usually sufficient to inspect the first 20 or so estimated autocorrelations of the original series, and of its first and second differences, if necessary.

Overdifferencing. Once stationarity is achieved, further differencing should be avoided. Overdifferencing introduces extra serial correlation and increases model complexity. To illustrate this point, assume that the series z_t follows a random walk so that the differenced series $w_t = (1 - B)z_t = a_t$ is white noise and thus stationary. Further differencing of w_t leads to $(1 - B)w_t = (1 - B)a_t$, which is a MA(1) model for w_t with parameter $\theta = 1$. Thus, the resulting model for z_t would be an ARIMA(0, 2, 1) model instead of the simpler ARIMA(0, 1, 0) model. The model with $\theta = 1$ is noninvertible and the pure autoregressive representation does not exit. Noninvertibility also causes problems at the parameter estimation stage in that approximate maximum likelihood methods tends to produce biased estimates in this case.

Figure 6.1 shows the autocorrelation and partial autocorrelation functions of a time series of length 200 generated from a random walk model with innovations variance equal to 1. The first 1000 observations were discarded to eliminate potential start-up effects. The estimated autocorrelations up to lag 20 of the original series and its first and second differences are shown in the graph. The autocorrelations of the original series fail to damp out quickly,

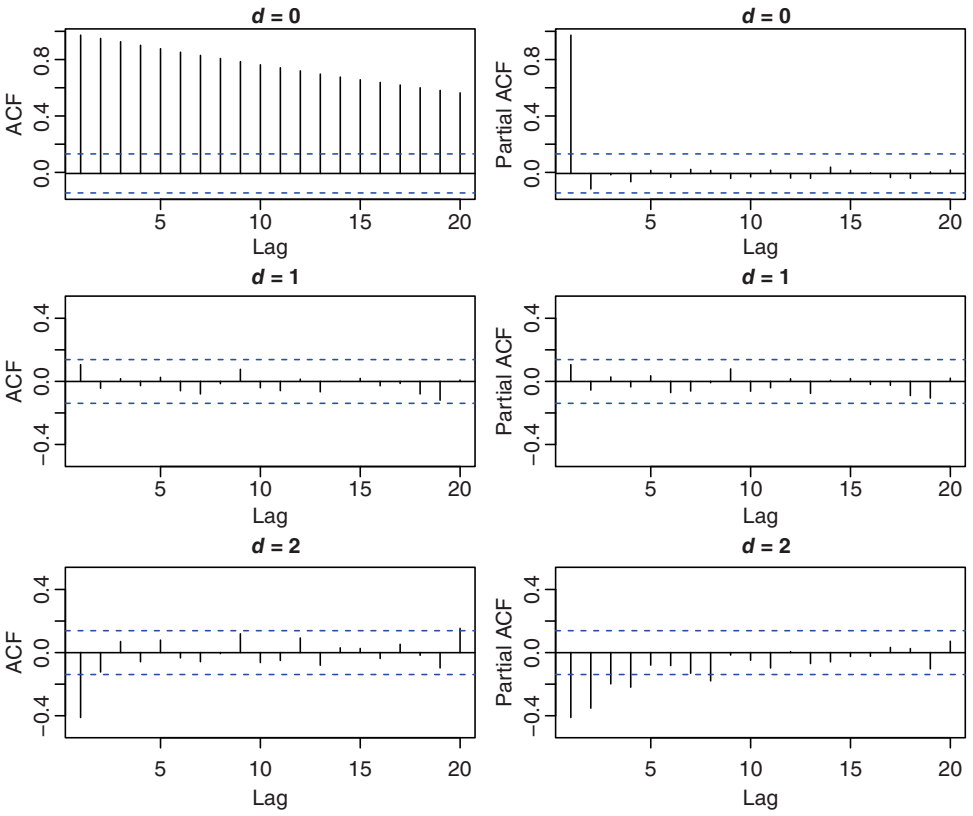

FIGURE 6.1 Estimated autocorrelation and partial autocorrelation functions for a simulated random walk process and its first ($d = 1$) and second ($d = 2$) differences.

indicating a need for differencing. The autocorrelations of $w_t = \nabla z_t$, on the other hand, are all small, demonstrating that stationarity has now been achieved. The autocorrelation function of the second differences $w_t = \nabla^2 z_t$ also indicates stationarity, but it has a spike at lag 1 showing the extra correlation that has emerged because of overdifferencing. The value of r_1 is close to -0.5, which is consistent with the lag 1 autocorrelation coefficient for an MA(1) model with $\theta = 1$. Figure 6.1 can be reproduced in R as follows:

```
> RW=arima.sim(list(order=c(0,1,0)),n=200,n.start=1000)
> acf0=acf(RW,20)
> pacf0=pacf(RW,20)
> acf1=acf(diff(RW),20)
> pacf1=pacf(diff(RW),20)
> acf2=acf(diff(diff(RW)),20)
> pacf2=pacf(diff(diff(RW)),20)
> par(mfrow=c(3,2))
> plot(acf0,main='d=0')
> plot(pacf0,main='d=0')
> plot(acf1,ylim=c(-0.5,0.5),main='d=1')
> plot(pacf1,ylim=c(-0.5,0.5),main='d=1')
```

```
> plot(acf2,ylim=c(-0.5,0.5),main='d=2')
> plot(pacf2,ylim=c(-0.5,0.5),main='d=2')
```

Identifying a Stationary ARMA Model for the Differenced Series. Having tentatively decided on the degree of differencing d, we examine the patterns of the estimated autocorrelation and partial autocorrelation functions of the differenced series, $w_t = (1 - B)^d z_t$, to determine a suitable choice for the orders p and q of the autoregressive and moving average operators. Here we recall the characteristic behavior of the theoretical autocorrelation and partial autocorrelation functions for moving average, autoregressive, and mixed processes, discussed in Chapter 3.

Briefly, whereas the autocorrelation function of an autoregressive process of order p tails off, its partial autocorrelation function has a cutoff after lag p. Conversely, the autocorrelation function of a moving average process of order q has a cutoff after lag q, while its partial autocorrelation function tails off. If both the autocorrelations and partial autocorrelations tail off, a mixed process is suggested. Furthermore, the autocorrelation function for a mixed process, containing a pth-order autoregressive component and a qth-order moving average component, is a mixture of exponentials and damped sine waves after the first $q-p$ lags. Conversely, the partial autocorrelation function for a mixed process is dominated by a mixture of exponentials and damped sine waves after the first $p-q$ lags (see Table 3.2).

In general, autoregressive (moving average) behavior, as measured by the autocorrelation function, tends to mimic moving average (autoregressive) behavior as measured by the partial autocorrelation function. For example, the autocorrelation function of a first-order autoregressive process decays exponentially, while the partial autocorrelation function cuts off after the first lag. Correspondingly, for a first-order moving average process, the autocorrelation function cuts off after the first lag. Although not precisely exponential, the partial autocorrelation function is dominated by exponential terms and has the general appearance of an exponential.

Of particular importance are the autoregressive and moving average processes of first and second order and the simple mixed $(1, d, 1)$ process. The properties of the theoretical autocorrelation and partial autocorrelation functions for these processes are summarized in Table 6.1, which requires careful study and provides a convenient reference table. The reader should also refer to Figures 3.2, 3.7, and 3.10, which show typical behavior of the autocorrelation function and the partial autocorrelation function for the second-order autoregressive process, the second-order moving average process, and the simple mixed ARMA(1, 1) process.

6.2.2 Standard Errors for Estimated Autocorrelations and Partial Autocorrelations

Estimated autocorrelations can have rather large variances and can be highly autocorrelated with each other. For this reason, *detailed* adherence to the theoretical autocorrelation function cannot be expected in the estimated function. In particular, moderately large estimated autocorrelations can occur after the theoretical autocorrelation function has damped out, and apparent ripples and trends can occur in the estimated function that have no basis in the theoretical function. In employing the estimated autocorrelation function as a tool for identification, it is usually possible to be fairly sure about broad characteristics, but more subtle indications may or may not represent real effects. For these reasons, two or more related

TABLE 6.1 Behavior of the Autocorrelation Functions for the dth Difference of an ARIMA Process of Order (p, d, q)[a]

	Order				
	$(1, d, 0)$	$(0, d, 1)$	$(2, d, 0)$	$(0, d, 2)$	$(1, d, 1)$
Behavior of ρ_k	Decays exponentially	Only ρ_1 nonzero	Mixture of exponentials or damped sine wave	Only ρ_1 and ρ_2 nonzero	Decays exponentially from first lag
Behavior of ϕ_{kk}	Only ϕ_{11} nonzero	Exponential dominates decay	Only ϕ_{11} and ϕ_{22} nonzero	Dominated by mixture of exponential or damped sine wave	Dominated by exponential decay from first lag
Preliminary estimates from	$\phi_1 = \rho_1$	$\rho_1 = \dfrac{-\theta_1}{1+\theta_1^2}$	$\phi_1 = \dfrac{\rho_1(1-\rho_2)}{1-\rho_1^2}$	$\rho_1 = \dfrac{-\theta_1(1-\theta_2)}{1+\theta_1^2+\theta_2^2}$	$\rho_1 = \dfrac{(1-\theta_1\phi_1)(\phi_1-\theta_1)}{1+\theta_1^2-2\phi_1\theta_1}$
			$\phi_2 = \dfrac{\rho_2-\rho_1^2}{1-\rho_1^2}$	$\rho_2 = \dfrac{-\theta_2}{1+\theta_1^2+\theta_2^2}$	$\rho_2 = \phi_1\rho_1$
Admissible region	$-1 < \phi_1 < 1$	$-1 < \theta_1 < 1$	$-1 < \phi_2 < 1$ $\phi_2 + \phi_1 < 1$ $\phi_2 - \phi_1 < 1$	$-1 < \theta_2 < 1$ $\theta_2 + \theta_1 < 1$ $\theta_2 - \theta_1 < 1$	$-1 < \phi_1 < 1$ $-1 < \theta_1 < 1$

[a] Table A and Charts B–D are included at the end of this book to facilitate the calculation of approximate estimates of the parameters for first-order moving average, second-order autoregressive, second-order moving average, and the mixed ARMA(1, 1) processes.

models may need to be entertained and investigated further at the estimation and diagnostic checking stages of model building.

In practice, it is important to have some indication of how far an estimated value may differ from the corresponding theoretical value. In particular, we need some means for judging whether the autocorrelations and partial autocorrelations are effectively zero after some specific lag q or p, respectively. For *larger lags*, on the hypothesis that the process is moving average of order q, we can compute standard errors of estimated autocorrelations from the simplified form of Bartlett's formula (2.1.15), with sample estimates replacing theoretical autocorrelations. Thus,

$$\hat{\sigma}[r_k] \simeq \frac{1}{n^{1/2}}[1 + 2(r_1^2 + r_2^2 + \cdots + r_q^2)]^{1/2} \qquad k > q \tag{6.2.2}$$

For the partial autocorrelations, we use the result quoted in (3.2.36) that, on the hypothesis that the process is autoregressive of order p, the standard error for estimated partial autocorrelations of order $p + 1$ and higher is

$$\hat{\sigma}[\hat{\phi}_{kk}] \simeq \frac{1}{n^{1/2}} \qquad k > p \tag{6.2.3}$$

It was shown by Anderson (1942) that for moderate n, the distribution of an estimated autocorrelation coefficient, whose theoretical value is zero, is approximately normal. Thus, on the hypothesis that the theoretical autocorrelation ρ_k is zero, the estimate r_k divided by its standard error will be approximately distributed as a unit normal variate. A similar result is true for the partial autocorrelations. These facts provide an informal guide as to whether theoretical autocorrelations and partial autocorrelations beyond a particular lag are essentially zero.

6.2.3 Identification of Models for Some Actual Time Series

Series A–D. In this section, the model specification tools described above are applied to some of the actual time series that we encountered in earlier chapters. We first discuss potential models for Series A to D plotted in Figure 4.1. As remarked in Chapter 4 on nonstationarity, we expect Series A, C, and D to possess nonstationary characteristics since they represent the ''uncontrolled'' behavior of certain process outputs. Similarly, we would expect the IBM stock price Series B to have no fixed level and to be nonstationary.

The estimated autocorrelations of z_t and the first differences ∇z_t for Series A–D are shown in Figure 6.2. Figure 6.3 shows the corresponding estimated partial autocorrelations. The two figures were generated in R using commands similar to those used to produce Figure 6.1. For the chemical process concentration readings in Series A, the autocorrelations for ∇z_t are small after the first lag. This suggests that this time series might be described by an IMA(0, 1, 1) model. However, from the autocorrelation function of z_t, it is seen that after lag 1 the correlations do decrease fairly regularly. Therefore, an alternative is that the series follows a mixed ARMA(1, 0, 1) model. The partial autocorrelation function of z_t seems to support this possibility. We will see later that the two alternatives result in virtually the same model. For the stock price Series B, the results confirm the nonstationarity of the original series and suggest that a random walk model $(1 - B)z_t = a_t$ is appropriate for this series.

The estimated autocorrelations of the temperature Series C also indicate nonstationarity. The roughly exponential decay in the autocorrelations for the first difference suggests a

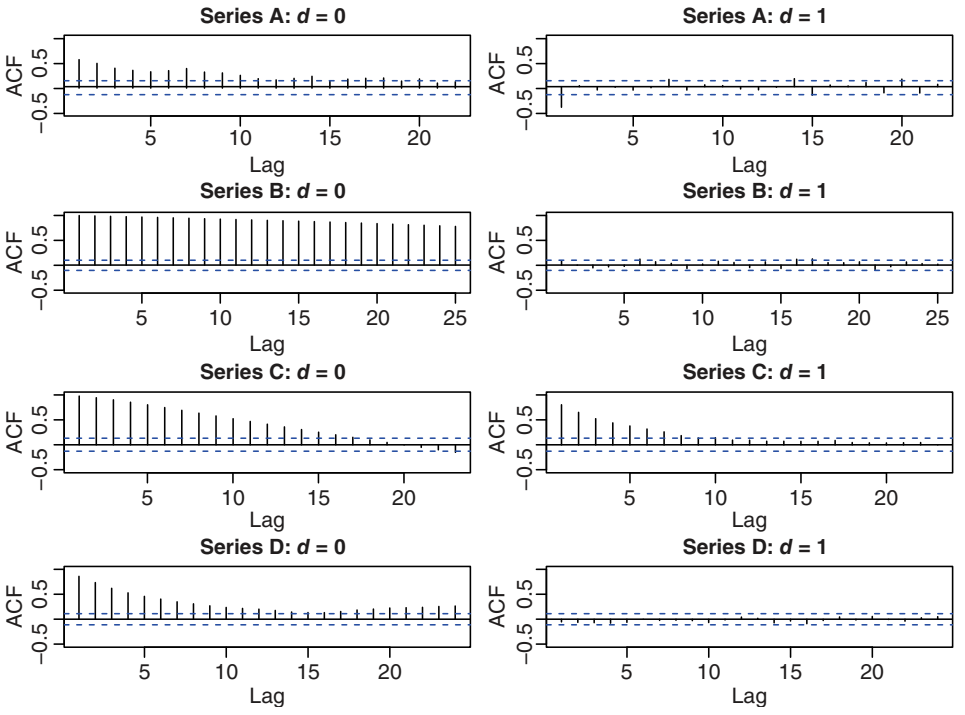

FIGURE 6.2 Estimated autocorrelation functions of the original series ($d = 0$), and their first differences ($d = 1$) for Series A–D.

process of order $(1, 1, 0)$, with an autoregressive parameter ϕ around 0.8. Alternatively, we notice that the autocorrelations of ∇z_t decay at a relatively slow rate, suggesting that further differencing might be needed. The autocorrelation and partial autocorrelation functions of the second differences $\nabla^2 z_t$ (not shown) were rather small, suggesting a white noise process for the second differences. This implies that an IMA$(0, 2, 0)$ model might also be appropriate for this series. Thus, the possibilities are

$$(1 - 0.8B)(1 - B)z_t = a_t$$
$$(1 - B)^2 z_t = a_t$$

The second model is very similar to the first, differing only in the choice of 0.8 rather than 1.0 for the autoregressive coefficient.

Finally, the autocorrelation and partial autocorrelation functions for the viscosity Series D suggest that an AR(1) model $(1 - \phi B)z_t = a_t$ with ϕ around 0.8 might be appropriate for this series. Alternatively, since the autocorrelation coefficients decay at a relatively slow rate, we will also consider the model $(1 - B)z_t = a_t$ for this series.

Series E and F. Series E shown in the top graph of Figure 6.4 represents the annual Wölfer sunspot numbers over the period 1770–1869. This series is likely to be stationary since the number of sunspots is expected to remain in equilibrium over long periods of time. The autocorrelation and partial autocorrelation functions in Figure 6.4 show characteristics similar to those of an AR(2) process. However, as will be seen later, a marginally better

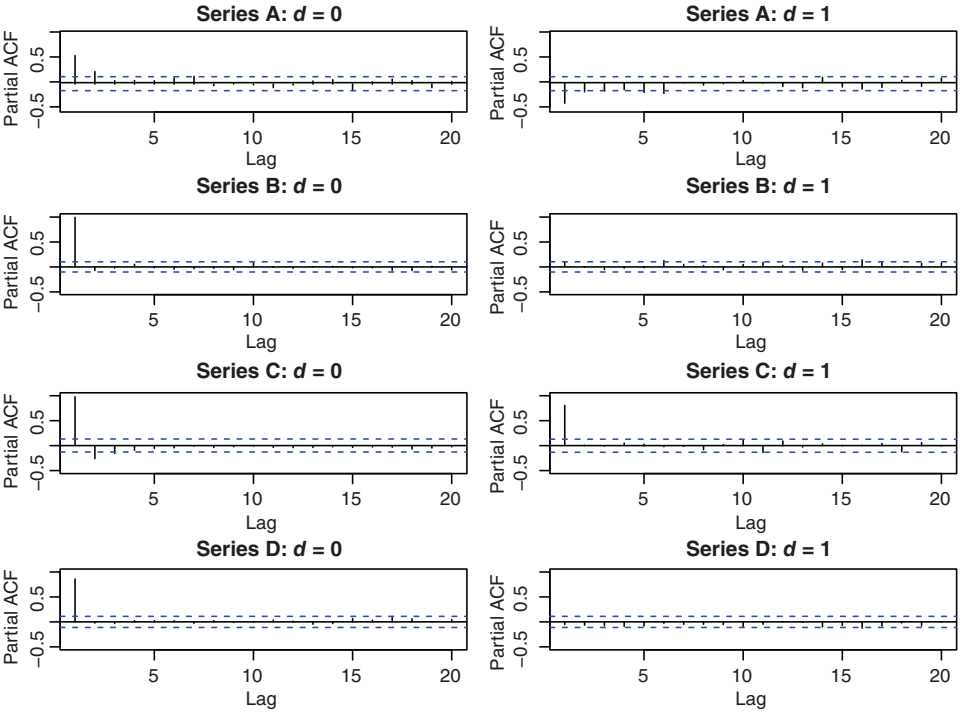

FIGURE 6.3 Estimated partial autocorrelation functions of the original series ($d = 0$), and their first differences ($d = 1$) for Series A–D.

fit is obtained using an AR(3) model. The fit can be improved further using a square root or log-transformation of the series. An autoregressive model of order nine is suggested by the order selection routine ar() in the R package that selects AR order based on the Akaike information criterion (AIC) to be discussed in Section 6.2.4. Other options considered in the literature include nonlinear time series models, such as bilinear or threshold autoregressive models, discussed briefly in Section 10.3.

Series F introduced in Chapter 2 represents the yields of a batch chemical process. The series is expected to be stationary since the batches are processed under uniformly controlled conditions. The stationarity is confirmed by Figure 6.5 that shows a graph of the series along with the autocorrelation and partial autocorrelation functions of the series and its first differences. The results for the undifferenced series suggest that a first-order autoregressive model might be appropriate for this series.

A summary of the models tentatively identified for Series A to F is given in Table 6.2. Note that for Series C and F, the alternative models suggested above have been made slightly more general for further illustrations later on.

Notes on the identification procedure. The graphs of the autocorrelation and partial autocorrelation functions shown above were generated using R. In assessing the estimated correlation functions, it is very helpful to plot one or two standard error limits around zero for the estimated coefficients. Limits from the R package are included in the graphs displayed above. These limits are approximate two standard error limits, $\pm 2/\sqrt{(n)}$, determined

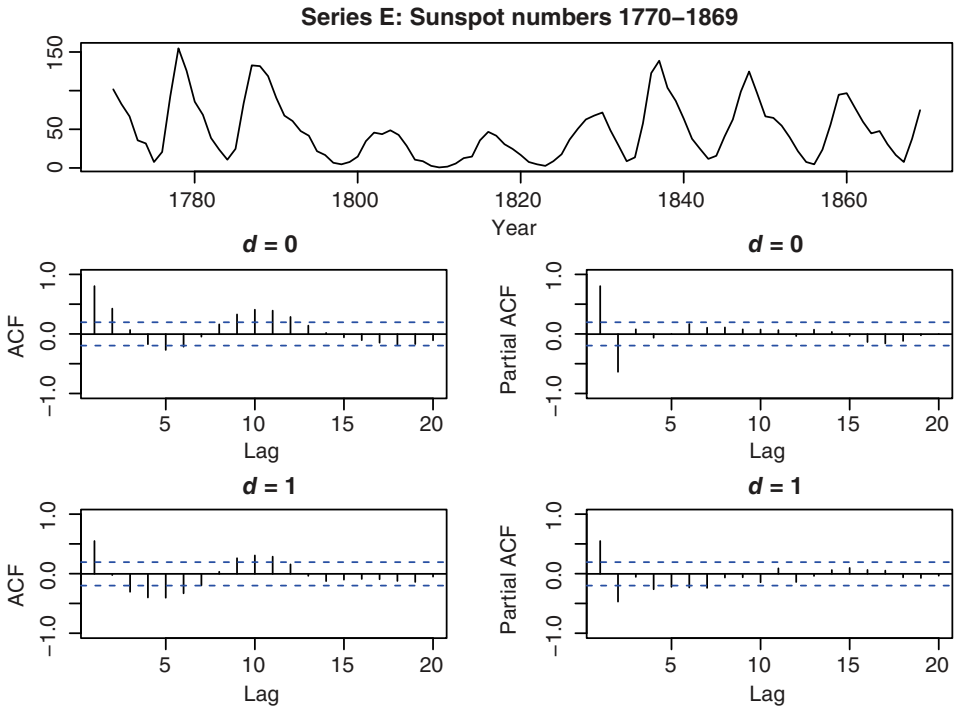

FIGURE 6.4 Estimated autocorrelation and partial autocorrelation functions of the sunspot series (Series E) and its first differences.

under the assumption that all the theoretical autocorrelation coefficients are zero so that the series is white noise. If a hypothesis about a specific model is postulated, alternative limits could be determined from Bartlett's formula as discussed above. When the calculations are performed in R, inclusion of the argument ci.type='ma' in the acf() function

TABLE 6.2 Tentative Identification of Models for Series A–F

Series	Degree of Differencing	Apparent Nature of Differenced Series	Identification for z_t
A	Either 0	Mixed first-order AR with first-order MA	(1, 0, 1)
	or 1	First-order MA	(0, 1, 1)
B	1	First-order MA	(0, 1, 1)
C	Either 1	First-order AR	(1, 1, 0)
	or 2	Uncorrelated noise	(0, 2, 2)[a]
D	Either 0	First-order AR	(1, 0, 0)
	or 1	Uncorrelated noise	(0, 1, 1)[a]
E	Either 0	Second-order AR	(2, 0, 0)
	or 0	Third-order AR	(3, 0, 0)
F	0	Second-order AR	(2, 0, 0)

[a] The order of the moving average operator appears to be zero, but the more general form is retained for subsequent consideration.

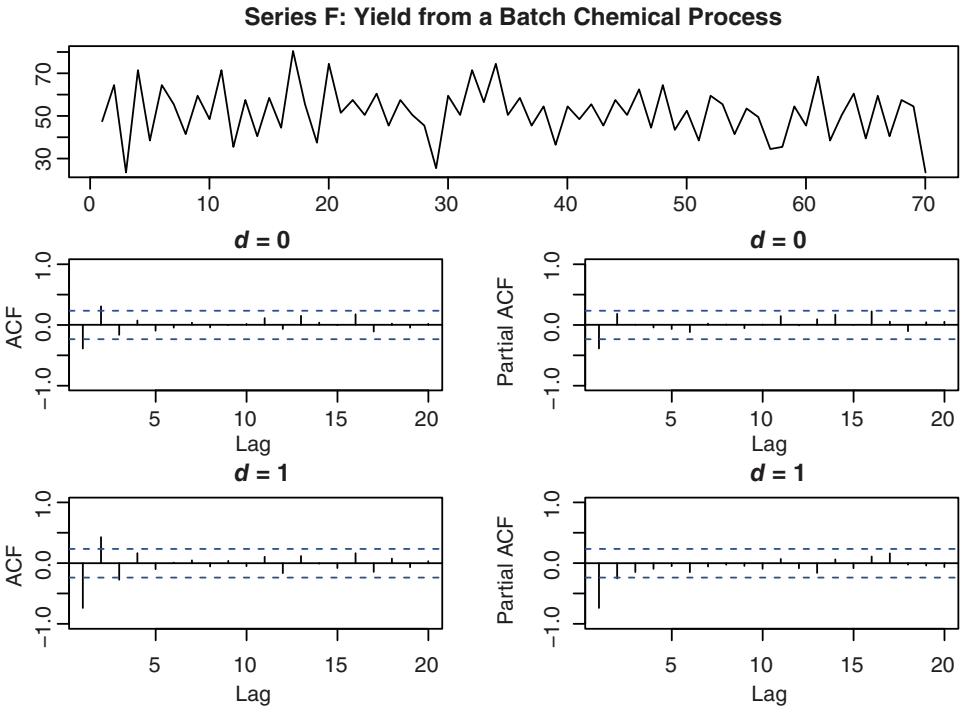

FIGURE 6.5 Estimated autocorrelation and partial autocorrelation functions for the yield of a batch chemical process (Series F) and its first differences.

yields confidence bounds computed based on the assumption that the true model is MA($k-1$).

Three other points concerning this identification procedure need to be mentioned:

1. Simple differencing of the kind we have used will not produce stationarity in series containing seasonal components. In Chapter 9, we discuss the appropriate modifications for such seasonal time series.
2. As discussed in Chapter 4, a nonzero value for θ_0 in (6.1.1) implies the existence of a systematic polynomial trend of degree d. For the nonstationary models in Table 6.2, a value of $\theta_0 = 0$ can perfectly well account for the behavior of the series. Occasionally, however, there will be some real physical phenomenon requiring the provision of such a component. In other cases, it might be uncertain whether or not such a provision should be made. Some indication of the evidence supplied by the data, for the inclusion of θ_0 in the model, can be obtained at the identification stage by comparing the mean $\overline{w}$ of $w_t = \nabla^d z_t$ with its approximate standard error, using $\sigma^2(\overline{w}) = n^{-1}\sigma_w^2[1 + 2\rho_1(w) + 2\rho_2(w) + \cdots]$.
3. It was noted in Section 3.4.2 that, for any ARMA(p, q) process with $p - q > 0$, the *whole positive half* of the autocorrelation function will be a mixture of damped sine waves and exponentials. This does not, of course, prevent us from tentatively identifying q, because (a) the partial autocorrelation function will show $p - q$ ''anomalous'' values before behaving like that of an MA(q) process, and (b) q must be such that the

autocorrelation function could take, as starting values following the general pattern, ρ_q back to $\rho_{-(p-q-1)}$.

6.2.4 Some Additional Model Identification Tools

Although the sample autocorrelation and partial autocorrelation functions are extremely useful in model identification, there are sometimes cases involving mixed models where they can provide ambiguous results. This may not be a serious problem since, as has been emphasized, model specification is always tentative and subject to further examination, diagnostic checking, and modification, if necessary. Nevertheless, there has been considerable interest in developing additional tools for use at the model identification stage. These include the R and S array approach proposed by Gray et al. (1978), the generalized partial autocorrelation function studied by Woodward and Gray (1981), the inverse autocorrelation function considered by Cleveland (1972) and Chatfield (1979), the extended sample autocorrelation function of Tsay and Tiao (1984), and the use of canonical correlation analysis as examined by Akaike (1976), Cooper and Wood (1982), and Tsay and Tiao (1985). Model selection criteria such as the AIC criterion introduced by Akaike (1974a) and the Bayesian Information Criterion (BIC) of Schwarz (1978) are also useful supplementary tools.

Canonical Correlation Methods. For illustration, we briefly discuss the use of canonical correlation analysis for model identification. In general, for two sets of random variables, $\boldsymbol{Y}_1 = (y_{11}, y_{12}, \ldots, y_{1k})'$ and $\boldsymbol{Y}_2 = (y_{21}, y_{22}, \ldots, y_{2l})'$, of dimensions k and l (assume $k \leq l$), canonical correlation analysis involves determining linear combinations $U_i = \boldsymbol{a}_i'\boldsymbol{Y}_1$ and $V_i = \boldsymbol{b}_i'\boldsymbol{Y}_2$, $i = 1, \ldots, k$, and corresponding correlations $\rho(i) = \text{corr}[U_i, V_i]$ with $\rho(1) \geq \rho(2) \geq \cdots \geq \rho(k) \geq 0$. The linear combinations are chosen so that the U_i and V_j are mutually uncorrelated for $i \neq j$, U_1 and V_1 have the maximum possible correlation $\rho(1)$ among all linear combinations of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, U_2 and V_2 have the maximum possible correlation $\rho(2)$ among all linear combinations of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ that are uncorrelated with U_1 and V_1, and so on. The resulting correlations $\rho(i)$ are called the *canonical correlations* between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, and the variables U_i and V_i are the corresponding canonical variates. If $\boldsymbol{\Omega} = \text{cov}[\boldsymbol{Y}]$ denotes the covariance matrix of $\boldsymbol{Y} = (\boldsymbol{Y}_1', \boldsymbol{Y}_2')'$, with $\boldsymbol{\Omega}_{ij} = \text{cov}[\boldsymbol{Y}_i, \boldsymbol{Y}_j]$, then it is known that the values $\rho^2(i)$ are the ordered eigenvalues of the matrix $\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}$ and the vectors $\boldsymbol{a}_i$, such that $U_i = \boldsymbol{a}_i'\boldsymbol{Y}_1$, are the corresponding (normalized) eigenvectors; that is, the $\rho^2(i)$ and $\boldsymbol{a}_i$ satisfy

$$[\rho^2(i)\mathbf{I} - \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}]\boldsymbol{a}_i = \mathbf{0} \qquad i = 1, \ldots, k \tag{6.2.4}$$

with $\rho^2(1) \geq \rho^2(2) \geq \cdots \geq \rho^2(k) \geq 0$ (e.g., Anderson (1984), p. 490). Similarly, one can define the notion of *partial canonical correlations* between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, given another set of variables $\boldsymbol{Y}_3$, as the canonical correlations between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ after they have been ''adjusted'' for the effects of $\boldsymbol{Y}_3$ by linear regression on Y_3, analogous to the definition of partial correlations as discussed in Section 3.2.5. A useful property to note is that if there exist (at least) $s \leq k$ linearly independent linear combinations of $\boldsymbol{Y}_1$ that are completely uncorrelated with $\boldsymbol{Y}_2$, say $\boldsymbol{U} = \mathbf{A}'\boldsymbol{Y}_1$ such that $\text{cov}[\boldsymbol{Y}_2, \boldsymbol{U}] = \boldsymbol{\Omega}_{21}\mathbf{A} = \mathbf{0}$, then there are (at least) s zero canonical correlations between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$. This follows easily from (6.2.4) since there will be (at least) s linearly independent eigenvectors satisfying (6.2.4) with

corresponding $\rho(i) = 0$. In effect, then, the number s of *zero* canonical correlations is equal to $s = k - r$, where $r = \text{rank}(\boldsymbol{\Omega}_{21})$.

In the ARMA time series model context, following the approach of Tsay and Tiao (1985), we consider $\boldsymbol{Y}_{m,t} = (\tilde{z}_t, \tilde{z}_{t-1}, \ldots, \tilde{z}_{t-m})'$ and examine the canonical correlation structure between the variables $\boldsymbol{Y}_{m,t}$ and

$$\boldsymbol{Y}_{m,t-j-1} = (\tilde{z}_{t-j-1}, \tilde{z}_{t-j-2}, \ldots, \tilde{z}_{t-j-1-m})'$$

for various combinations of $m = 0, 1, \ldots$ and $j = 0, 1, \ldots$ A key feature to recall is that the autocovariance function γ_k of an ARMA(p, q) process $\tilde{z}_t$ satisfies (3.4.2), and, in particular,

$$\gamma_k - \sum_{i=1}^{p} \phi_i \gamma_{k-i} = 0 \qquad k > q$$

Thus, for example, if $m \geq p$, there is (at least) one linear combination of $\boldsymbol{Y}_{m,t}$,

$$\tilde{z}_t - \sum_{i=1}^{p} \phi_i \tilde{z}_{t-1} = (1, -\phi_1, \ldots, -\phi_p, 0, \ldots, 0)\boldsymbol{Y}_{m,t} = \boldsymbol{a}'\boldsymbol{Y}_{m,t} \tag{6.2.5}$$

such that

$$\boldsymbol{a}'\boldsymbol{Y}_{m,t} = a_t - \sum_{i=1}^{q} \theta_i a_{t-i}$$

which is uncorrelated with $\boldsymbol{Y}_{m,t-j-1}$ for $j \geq q$. In particular, then, for $m = p$ and $j = q$, there is one zero canonical correlation between $\boldsymbol{Y}_{p,t}$ and $\boldsymbol{Y}_{p,t-q-1}$, as well as between $\boldsymbol{Y}_{p,t}$ and $\boldsymbol{Y}_{p,t-j-1}$, $j > q$, and between $\boldsymbol{Y}_{m,t}$ and $\boldsymbol{Y}_{m,t-q-1}$, $m > p$, while in general it is not difficult to establish that there are $s = \min(m + 1 - p, j + 1 - q)$ zero canonical correlations between $\boldsymbol{Y}_{m,t}$ and $\boldsymbol{Y}_{m,t-j-1}$ for $m > p$ and $j > q$. Hence, one can see that determination of the structure of the zero canonical correlations between $\boldsymbol{Y}_{m,t}$ and $\boldsymbol{Y}_{m,t-j-1}$ for various values of m and j will serve to characterize the orders p and q of the ARMA model, and so the canonical correlations will be useful in model identification. We note the special cases of these canonical correlations are as follows. First, when $m = 0$, we are simply examining the autocorrelations ρ_{j+1} between z_t and z_{t-j-1}, which will all equal zero in an MA(q) process for $j \geq q$. Second, when $j = 0$, we are examining the partial autocorrelations $\phi_{m+1,m+1}$ between z_t and z_{t-m-1}, given $z_{t-1}, \ldots, z_{t-m}$, and these will all equal zero in an AR(p) process for $m \geq p$. Hence, the canonical correlation analysis can be viewed as an extension of the analysis of the autocorrelation and partial autocorrelation functions of the process.

In practice, based on (6.2.4), one is led to consider the sample canonical correlations $\hat{\rho}(i)$, which are determined from the eigenvalues of the matrix:

$$\left(\sum\nolimits_t \boldsymbol{Y}_{m,t}\boldsymbol{Y}'_{m,t}\right)^{-1} \left(\sum\nolimits_t \boldsymbol{Y}_{m,t}\boldsymbol{Y}'_{m,t-j-1}\right) \times \left(\sum\nolimits_t \boldsymbol{Y}_{m,t-j-1}\boldsymbol{Y}'_{m,t-j-1}\right)^{-1} \left(\sum\nolimits_t \boldsymbol{Y}_{m,t-j-1}\boldsymbol{Y}'_{m,t}\right) \tag{6.2.6}$$

for various values of lag $j = 0, 1, \ldots$ and $m = 0, 1, \ldots$. Tsay and Tiao (1985) use a chi-squared test statistic approach based on the *smallest* eigenvalue (squared sample canonical correlation) $\hat{\lambda}(m, j)$ of (6.2.6). They propose the statistic $c(m, j) = -(n - m - j)\ln[1 -$

$\hat{\lambda}(m,j)/d(m,j)]$, where $d(m,j) = 1 + 2\sum_{i=1}^{j} r_i^2(w'), j > 0, r_i(w')$ denotes the sample autocorrelation at lag i of $w_t' = z_t - \hat{\phi}_1^{(j)} z_{t-1} - \cdots - \hat{\phi}_m^{(j)} z_{t-m}$, and the $\hat{\phi}_i^{(j)}$ are estimates of the ϕ_i's obtained from the eigenvector (see, for example, equation (6.2.5)) corresponding to $\hat{\lambda}(m,j)$. The statistic $c(m,j)$ has an asymptotic χ_1^2 distribution when $m = p$ and $j \geq q$ or when $m \geq p$ and $j = q$ and can be used to test whether there exists a zero canonical correlation in theory. Hence if the sample statistics exhibit a pattern such that they are all insignificant, relative to a χ_1^2 distribution, for $m \geq p$ and $j \geq q$ for some p and q values, then the model might reasonably be identified as an ARMA(p,q) for the smallest values (p,q) such that this pattern holds. Tsay and Tiao (1985) also show that this procedure is valid for nonstationary ARIMA models $\varphi(B)z_t = \theta(B)a_t$, in the sense that the overall order $p + d$ of the generalized AR operator $\varphi(B)$ can be determined by the procedure, without initially deciding on differencing of the original series z_t.

Canonical correlation methods were previously also proposed for ARMA modeling by Akaike (1976) and Cooper and Wood (1982). Their approach is to perform a canonical correlation analysis between the vector of present and past values, $\boldsymbol{P}_t \equiv \boldsymbol{Y}_{m,t} = (\tilde{z}_t, \tilde{z}_{t-1}, \ldots, \tilde{z}_{t-m})'$, and the vector of future values, $\boldsymbol{F}_{t+1} = (\tilde{z}_{t+1}, \tilde{z}_{t+2}, \ldots)'$. In practice, the finite lag m used to construct the vector of present and past values $\boldsymbol{P}_t$ may be fixed by use of an order determination criterion such as Akaike information criteria to be discussed a little later in this section, applied to fitting of AR models of various orders. The canonical correlation analysis is performed sequentially by adding elements to $\boldsymbol{F}_{t+1}$ one at a time, starting with $\boldsymbol{F}_{t+1}^* = (\tilde{z}_{t+1})$, until the first zero canonical correlation between $\boldsymbol{P}_t$ and the $\boldsymbol{F}_{t+1}$ is determined. Akaike (1976) uses an AIC-type criterion called deviance information criterion (DIC) to judge whether the smallest sample canonical correlation can be taken to be zero, while Cooper and Wood (1982) use a traditional chi-squared statistic approach to assess the significance of the smallest canonical correlation, although as pointed out by Tsay (1989a), to be valid in the presence of a moving average component, this statistic needs to be modified.

At a given stage in the procedure, when the smallest sample canonical correlation between $\boldsymbol{P}_t$ and $\boldsymbol{F}_{t+1}^*$ is first judged to be 0 and $\tilde{z}_{t+K+1}$ is the most recent variable to be included in $\boldsymbol{F}_{t+1}^*$, a linear combination of $\tilde{z}_{t+K+1}$ in terms of the remaining elements of $\boldsymbol{F}_{t+1}^*$ is identified that is uncorrelated with the past. Specifically, the linear combination $\tilde{z}_{t+K+1} - \sum_{j=1}^{K} \phi_j \tilde{z}_{t+K+1-j}$ of the elements in the vector $\boldsymbol{F}_{t+1}^*$ of future values is (in theory) determined to be uncorrelated with the past $\boldsymbol{P}_t$. Hence, this canonical correlation analysis procedure determines that the forecasts $\hat{z}_t(K+1)$ of the process satisfy

$$\hat{z}_t(K+1) - \sum_{j=1}^{K} \phi_j \hat{z}_t(K+1-j) = \theta_0$$

By reference to the relation (5.3.2) in Section 5.3, for a stationary process, this implies that an ARMA model is identified for the process, with $K = \max\{p,q\}$.

As can be seen, in the notation of Tsay and Tiao (1985), the methods of Akaike and Cooper and Wood represent canonical correlation analysis between $\boldsymbol{Y}_{m,t}$ and $\boldsymbol{Y}_{n-1,t+n}$ for various $n = 1, 2, \ldots$. Since the Tsay and Tiao method considers canonical correlation analysis between $\boldsymbol{Y}_{m,t}$ and $\boldsymbol{Y}_{m,t-j-1}$ for various combinations of $m = 0, 1, \ldots$ and $j = 0, 1, \ldots$, it is more general and, in principle, it is capable of providing information on the orders p and q of the AR and MA parts of the model separately, rather than just the maximum of these two values. In practice, when using the methods of Akaike and Cooper and Wood,

the more detailed information on the individual orders p and q would be determined at the stage of maximum likelihood estimation of the parameters of the ARMA(K, K) model.

Use of Model Selection Criteria. Another approach to model selection involves the use of information criteria such as AIC proposed by Akaike (1974a) or the Bayesian information criteria of Schwarz (1978). In the implementation of this approach, a range of potential ARMA models are estimated by maximum likelihood methods to be discussed in Chapter 7, and for each model, a criterion such as AIC (normalized by sample size n), given by

$$\text{AIC}_{p,q} = \frac{-2\ln(\text{maximized likelihood}) + 2r}{n} \approx \ln(\hat{\sigma}_a^2) + r\frac{2}{n} + \text{ constant}$$

or the related BIC given by

$$\text{BIC}_{p,q} = \ln(\hat{\sigma}_a^2) + r\frac{\ln(n)}{n}$$

is evaluated. Here, $\hat{\sigma}_a^2$ is the maximum likelihood estimate of σ_a^2, and $r = p + q + 1$ is the number of estimated parameters, including a constant term. In the above criteria, the first term essentially corresponds to $-2/n$ times the log of the maximized likelihood, while the second term is a ''penalty factor'' for inclusion of additional parameters in the model. In the information criteria approach, models that yield a minimum value for the criterion are to be preferred, and the AIC or BIC values are compared among various models as the basis for selection of the model. Hence, since the BIC criterion imposes a greater penalty for the number of estimated model parameters than does AIC, use of minimum BIC for model selection would always result in a chosen model whose number of parameters is no greater than that chosen under AIC.

Hannan and Rissanen (1982) proposed a two-step model selection procedure that avoids the need to maximize the likelihood function for multiple combinations of p and q. At the first step, one fits an AR model of sufficiently high order m^* to the series $\tilde{z}_t$. The residuals $\tilde{a}_t$ from the fitted AR(m^*) model provide estimates of the innovations a_t in the ARMA(p, q) model. At the second step, one regresses $\tilde{z}_t$ on $\tilde{z}_{t-1}, \ldots, \tilde{z}_{t-p}$ and $\tilde{a}_{t-1}, \ldots, \tilde{a}_{t-q}$, for various combinations of p and q. That is, one fits approximate models of the form

$$\tilde{z}_t = \sum_{j=1}^{p} \phi_j \tilde{z}_{t-j} - \sum_{j=1}^{q} \theta_j \tilde{a}_{t-j} + a_t \tag{6.2.7}$$

using ordinary least squares, and the estimated error variance, uncorrected for degrees of freedom, is denoted by $\hat{\sigma}_{p,q}^2$. Then, using the BIC criterion, the order (p, q) of the ARMA model is chosen as the one that minimizes $\ln(\hat{\sigma}_{p,q}^2) + (p + q)\ln(n)/n$. Hannnan and Rissanen show that, under very general conditions, the estimators of p and q chosen in this manner tend almost surely to the true values. The appeal of this procedure is that computation of maximum likelihood estimates over a wide range of possible ARMA models is avoided.

While these order selection procedures are useful, they should be viewed as supplementary tools to assist in the model selection process. In particular, they should not be used as a substitute for careful examination of the estimated autocorrelation and partial autocorrelation functions of the series, and critical examination of the residuals $\hat{a}_t$ from

a fitted model should always be included as a major part of the overall model selection process.

6.3 INITIAL ESTIMATES FOR THE PARAMETERS

6.3.1 Uniqueness of Estimates Obtained from the Autocovariance Function

While a given ARMA model has a unique autocovariance structure, the converse is not true without additional conventions imposed for uniqueness, as we discuss subsequently. At first sight this would seem to rule out the use of the estimated autocovariances as a means of identification. However, we show in Section 6.4 that the estimated autocovariance function may indeed be used for this purpose. The reason is that, although there exists a multiplicity of ARMA models possessing the same autocovariance function, there exists only one that expresses the current value of $w_t = \nabla^d z_t$, exclusively in terms of previous history and in stationary invertible form.

6.3.2 Initial Estimates for Moving Average Processes

As shown in Chapter 3, the first q autocorrelations of a MA(q) process are nonzero and can be written in terms of the parameters of the model as

$$\rho_k = \frac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2} \qquad k = 1, 2, \ldots, q \tag{6.3.1}$$

The expression (6.3.1) for $\rho_1, \rho_2, \ldots, \rho_q$, in terms of $\theta_1, \theta_2, \ldots, \theta_q$, supplies q equations in q unknowns. Preliminary estimates of the θ's can be obtained by substituting the estimates r_k for ρ_k in (6.3.1) and solving the resulting nonlinear equations. A preliminary estimate of σ_a^2 may then be obtained from

$$\gamma_0 = \sigma_a^2(1 + \theta_1^2 + \cdots + \theta_q^2)$$

by substituting the preliminary estimates of the θ's and replacing $\gamma_0 = \sigma_w^2$ by its estimate c_0. The numerical values of the estimated autocorrelation coefficients r_k for the series `z` are conveniently obtained from R as follows:

```
> ac=acf(z)
> ac
```

Preliminary Estimates for a (0, d, 1) Process. Table A in Part Five relates ρ_1 to θ_1, and by substituting $r_1(w)$ for ρ_1 can be used to provide initial estimates for any $(0, d, 1)$ process $w_t = (1 - \theta_1 B)a_t$, where $w_t = \nabla^d z_t$.

Preliminary Estimates for a (0, d, 2) Process. Chart C in Part Five relates ρ_1 and ρ_2 to θ_1 and θ_2, and by substituting $r_1(w)$ and $r_2(w)$ for ρ_1 and ρ_2 can be used to provide initial estimates for any $(0, d, 2)$ process.

In obtaining preliminary estimates in this way, the following points should be kept in mind:

1. The autocovariances are second moments of the joint distribution of the w's. Thus, the parameter estimates are obtained by equating sample moments to their theoretical values. It is well known that the *method of moments* is not necessarily efficient and can produce poor estimates for models that include moving average terms. However, the rough estimates obtained can be useful in obtaining fully efficient estimates, because they supply an approximate idea of ''where in the parameter space to look'' for the most efficient estimates.
2. In general, the equation (6.3.1), obtained by equating moments, will have multiple solutions. For instance, when $q = 1$,

$$\rho_1 = \frac{-\theta_1}{1 + \theta_1^2} \tag{6.3.2}$$

and hence from $\theta_1^2 + (1/\rho_1)\theta_1 + 1 = 0$, we see that both

$$\theta_1 = -\frac{1}{2\rho_1} + \left[\frac{1}{(2\rho_1)^2} - 1\right]^{1/2}$$

and

$$\theta_1 = -\frac{1}{2\rho_1} - \left[\frac{1}{(2\rho_1)^2} - 1\right]^{1/2} \tag{6.3.3}$$

are possible solutions. For illustration, the first lag autocorrelation of the first difference of Series A is about -0.4. Substitution in (6.3.3) yields the pair of solutions $\theta_1 \simeq 0.5$ and $\theta_1' \simeq 2.0$. However, the chosen value $\theta_1 \simeq 0.5$ is the only value that lies within the invertibility interval $-1 < \theta_1 < 1$. In fact, it is shown in Section 6.4.1 that it is always true that only one of the multiple solutions of (6.3.1) can satisfy the invertibility condition.

Examples. Series A, B, and D were all identified in Table 6.2 as possible IMA processes of order (0, 1, 1). We have seen in Section 4.3.1 that this model may be written in following the alternative forms:

$$\begin{aligned}
\nabla z_t &= (1 - \theta_1 B)a_t \\
\nabla z_t &= \lambda_0 a_{t-1} + \nabla a_t \qquad (\lambda_0 = 1 - \theta_1) \\
z_t &= \lambda_0 \sum_{j=1}^{\infty} (1 - \lambda_0)^{j-1} z_{t-j} + a_t
\end{aligned}$$

Using Table A in Part Five, the approximate estimates of the parameters shown in Table 6.3 were obtained.

Series C has been tentatively specified in Table 6.2 as an IMA(0, 2, 2) process:

$$\nabla^2 z_t = (1 - \theta_1 B - \theta_2 B^2)a_t$$

or equivalently,

$$\nabla^2 z_t = (\lambda_0 \nabla + \lambda_1)a_{t-1} + \nabla^2 a_t$$

TABLE 6.3 Initial Estimates of Parameters for Series A, B, and D

Series	r_1	$\hat{\theta}_1$	$\hat{\lambda}_0 = 1 - \hat{\theta}_1$
A	−0.41	0.5	0.5
B	0.09	−0.1	1.1
D	−0.05	0.1	0.9

Since the first two sample autocorrelations of $\nabla^2 z_t$ are very close to zero, Chart C in Part Five gives $\hat{\theta}_1 = 0, \hat{\theta}_2 = 0$, so that $\hat{\lambda}_0 = 1 + \hat{\theta}_2 = 1$ and $\hat{\lambda}_1 = 1 - \hat{\theta}_1 - \hat{\theta}_2 = 1$. On this basis, the series would be represented by

$$\nabla^2 Z_t = a_t \tag{6.3.4}$$

This would mean that the second difference, $\nabla^2 z_t$, was very nearly a random (white noise) series.

6.3.3 Initial Estimates for Autoregressive Processes

For an assumed AR process of order 1 or 2, initial estimates for ϕ_1 and ϕ_2 can be calculated by substituting estimates r_j for the theoretical autocorrelations ρ_j in the formulas of Table 6.1, which are obtained from the Yule–Walker equations (3.2.6). In particular, for an AR(1), $\hat{\phi}_{11} = r_1$, and for an AR(2),

$$\begin{aligned} \hat{\phi}_{21} &= \frac{r_1(1 - r_2)}{1 - r_1^2} \\ \hat{\phi}_{22} &= \frac{r_2 - r_1^2}{1 - r_1^2} \end{aligned} \tag{6.3.5}$$

where $\hat{\phi}_{pj}$ denotes the estimated jth autoregressive parameter in a process of order p. The corresponding formulas given by the Yule–Walker equations for higher order schemes may be obtained by substituting the r_j for the ρ_j in (3.2.7). Thus,

$$\hat{\boldsymbol{\phi}} = \mathbf{R}_p^{-1} \mathbf{r}_p \tag{6.3.6}$$

where $\mathbf{R}_p$ is an estimate of the $p \times p$ matrix $\mathbf{P}_p$, as depicted following (3.2.6) in 3.2.2, of autocorrelations up to order $p - 1$, and $\mathbf{r}_p = (r_1, r_2, \ldots, r_p)'$. For example, if $p = 3$, (6.3.6.) becomes

$$\begin{bmatrix} \hat{\phi}_{31} \\ \hat{\phi}_{32} \\ \hat{\phi}_{33} \end{bmatrix} = \begin{bmatrix} 1 & r_1 & r_2 \\ r_1 & 1 & r_1 \\ r_2 & r_1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix} \tag{6.3.7}$$

A simple recursive method due to Levinson and Durbin for obtaining the estimates for an AR(p) from those of an AR($p - 1$) was discussed in Appendix A3.2.

It will be shown in Chapter 7 that in contrast to the situation for MA processes, the autoregressive parameters obtained from (6.3.6) approximate the fully efficient maximum likelihood estimates.

Example. Series E representing the sunspot data behaves in its undifferenced form like[1] an autoregressive process of second order:

$$(1 - \phi_1 B - \phi_2 B^2)\tilde{z}_t = a_t$$

Substituting the estimates $r_1 = 0.81$ and $r_2 = 0.43$, obtained using R, into (6.3.5), we have $\hat{\phi}_1 = 1.32$ and $\hat{\phi}_2 = -0.63$.

As a second example, consider again Series C identified as either of order $(1, 1, 0)$ or possibly $(0, 2, 2)$. The first possibility would give

$$(1 - \phi_1 B)\nabla Z_t = a_t$$

with $\hat{\phi}_1 = 0.81$, since r_1 for ∇z_t is 0.81.

This example is interesting because it makes clear that the two alternative models that have been identified for this series are closely related. On the supposition that the series is of order $(0, 2, 2)$, we found in (6.3.4) that this simplifies to

$$(1 - B)(1 - B)z_t = a_t \tag{6.3.8}$$

The alternative

$$(1 - 0.81B)(1 - B)z_t = a_t \tag{6.3.9}$$

is very similar.

6.3.4 Initial Estimates for Mixed Autoregressive–Moving Average Processes

It is often found, either initially or after suitable differencing, that $w_t = \nabla^d z_t$ is most economically represented by a mixed ARMA process:

$$\phi(B)w_t = \theta(B)a_t$$

As noted in Section 6.2.1, a mixed process is indicated if both the autocorrelation and partial autocorrelation functions tail off rather than either having a cutoff feature. Another helpful fact in identifying the mixed process is that after lag $q - p$, the theoretical autocorrelations of the mixed process behave like the autocorrelations of a pure autoregressive process $\phi(B)w_t = a_t$ (see (3.4.3)). In particular, if the autocorrelation function of the dth difference appears to be falling off exponentially from an aberrant first value r_1, we would suspect that we have a process of order $(1, d, 1)$ that is,

$$(1 - \phi_1 B)w_t = (1 - \theta_1 B)a_t \tag{6.3.10}$$

where $w_t = \nabla^d z_t$.

[1]The sunspot series has been the subject of much investigation. Early references include Schuster (1906), Yule (1927), and Moran (1954). The series does not appear to be adequately represented by a second-order autoregressive process. A model related to the underlying mechanism at work would, of course, be the most satisfactory. More recent work has suggested empirically that a second-order autoregressive model would provide a better fit if a suitable transformation such as log or square root were first applied to z. Inclusion of a higher order term, at lag 9, in the AR model also improves the fit. Other possibilities include the use of nonlinear time series models, such as bilinear or threshold autoregressive models (e.g., see Section 10.3), as has been investigated by Subba Rao and Gabr (1984), Tong and Lim (1980), and Tong (1983,1990).

Approximate values for the parameters of the process (6.3.10) are obtained by substituting the estimates $r_1(w)$ and $r_2(w)$ for ρ_1 and ρ_2 in the expression (3.4.8). This gives

$$r_1 = \frac{(1 - \hat{\phi}_1\hat{\theta}_1)(\hat{\phi}_1 - \hat{\theta}_1)}{1 + \hat{\theta}_1^2 - 2\hat{\phi}_1\hat{\theta}_1}$$

$$r_2 = r_1\hat{\phi}_1$$

Chart D in Part Five relates ρ_1 and ρ_2 to ϕ_1 and θ_1 can be used to provide initial estimates of the parameters for any $(1, d, 1)$ process.

For example, using Figure 6.2, Series A was identified as of order $(0, 1, 1)$, with θ_1 about 0.5. Looking at the autocorrelation function of z_t rather than that of $w_t = \nabla z_t$, we see from r_1 onward the autocorrelations decay roughly exponentially, although slowly. Thus, an alternative specification for Series A is that it is generated by a stationary process of order $(1, 0, 1)$. The estimated autocorrelations and the corresponding initial estimates of the parameters are then

$$r_1 = 0.57 \qquad r_2 = 0.50 \qquad \hat{\phi}_1 \simeq 0.87 \qquad \hat{\theta}_1 \simeq 0.48$$

This identification yields the approximate model of order $(1, 0, 1)$:

$$(1 - 0.9B)\tilde{z}_t = (1 - 0.5B)a_t$$

whereas the previously identified model of order $(0, 1, 1)$, given in Table 6.5, is

$$(1 - B)z_t = (1 - 0.5B)a_t$$

Again we see that the ''alternative'' models are nearly the same.

Compensation between Autoregressive and Moving Average Operators. The alternative models identified above are even more alike than they appear. This is because small changes in the autoregressive operator of a mixed model can be nearly compensated by corresponding changes in the moving average operator. In particular, if we have a model

$$[1 - (1 - \delta)B]\tilde{z}_t = (1 - \theta B)a_t$$

where δ is small and positive, we can write

$$\begin{aligned}(1 - B)\tilde{z}_t &= [1 - (1 - \delta)B]^{-1}(1 - B)(1 - \theta B)a_t \\ &= \{1 - \delta B[1 + (1 - \delta)B + (1 - \delta)^2B^2 + \cdots]\}(1 - \theta B)a_t \\ &= [1 - (\theta + \delta)B]a_t + \text{terms in } a_{t-2}, a_{t-3}, \cdots, \text{of order } \delta\end{aligned}$$

6.3.5 Initial Estimate of Error Variance

For comparison with the more efficient methods of estimation to be described in Chapter 7, it is interesting to see how much additional information about the model can be extracted at the identification stage. We have already shown how to obtain initial estimates $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$ of the parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$ in the ARMA model, identified for an appropriate difference $w_t = \nabla^d z_t$ of the series. In this section we show how to obtain preliminary estimates of the error variance σ_a^2, and in Section 6.3.6 we show how to obtain an approximate standard error for the sample mean $\overline{w}$ of the appropriately differenced series.

An initial estimate of the error variance may be obtained by substituting an estimate c_0 in the expression for the variance γ_0 given in Chapter 3. Thus, substituting in (3.2.8), an initial estimate of σ_a^2 for an AR process may be obtained from

$$\hat{\sigma}_a^2 = c_0(1 - \hat{\phi}_1 r_1 - \hat{\phi}_2 r_2 - \cdots - \hat{\phi}_p r_p) \tag{6.3.11}$$

Similarly, from (3.3.3), an initial estimate for a MA process may be obtained from

$$\hat{\sigma}_a^2 = \frac{c_0}{1 + \hat{\theta}_1^2 + \cdots + \hat{\theta}_q^2} \tag{6.3.12}$$

The form of the estimate for a mixed process is, in general, more complicated. However, for the important ARMA(1,1) process, it takes the form (see (3.4.7))

$$\hat{\sigma}_a^2 = \frac{1 - \hat{\phi}_1^2}{1 + \hat{\theta}_1^2 - 2\hat{\phi}_1\hat{\theta}_1} c_0 \tag{6.3.13}$$

For example, consider the (1, 0, 1) model identified for Series A. Using (6.3.13) with $\hat{\phi}_1 = 0.87$, $\hat{\theta}_1 = 0.48$, and $c_0 = 0.1586$, we obtain the estimate $\hat{\sigma}_a^2 = 0.098$.

6.3.6 Approximate Standard Error for $\overline{w}$

The general ARIMA model, for which the mean μ_w of $w_t = \nabla^d z_t$ is not necessarily zero, may be written in any one of the three forms:

$$\phi(B)(w_t - \mu_w) = \theta(B)a_t \tag{6.3.14}$$

$$\phi(B)w_t = \theta_0 + \theta(B)a_t \tag{6.3.15}$$

$$\theta(B)w_t = \theta(B)(a_t + \xi) \tag{6.3.16}$$

where

$$\mu_w = \frac{\theta_0}{1 - \phi_1 - \phi_2 - \cdots - \phi_p} = \frac{(1 - \theta_1 - \theta_2 - \cdots - \theta_q)\xi}{1 - \phi_1 - \phi_2 - \cdots - \phi_p}$$

Hence, if $1 - \phi_1 - \phi_2 - \cdots - \phi_p \neq 0$ and $1 - \theta_1 - \theta_2 - \cdots - \theta_p \neq 0$, $\mu_w = 0$ implies that $\theta_0 = 0$ and $\xi = 0$. Now, in general, when $d = 0$, μ_z will not be zero. However, consider the eventual forecast function associated with the general model (6.3.14) when $d > 0$. With $\mu_w = 0$, this forecast function already contains an adaptive polynomial component of degree $d - 1$. The effect of allowing μ_w to be nonzero is to introduce a *fixed* polynomial term into this function of degree d. For example, if $d = 2$ and μ_w is nonzero, the forecast function $\hat{z}_t(l)$ includes a quadratic component in l, in which the coefficient of the quadratic term is fixed and does not adapt to the series. Because models of this kind are often inapplicable when $d > 0$, the hypothesis that $\mu_w = 0$ will frequently not be contradicted by the data. Indeed, as we have indicated, we usually assume that $\mu_w = 0$ unless evidence to the contrary presents itself.

At this, the identification stage of model building, an indication of whether or not a nonzero value for μ_w is needed may be obtained by comparison of $\overline{w} = \sum_{t=1}^{n} w_t/n$ with

its approximate standard error (see Section 2.1.5). With $n = N - d$ differences available,

$$\sigma^2(\overline{w}) = n^{-1}\gamma_0 \sum_{-\infty}^{\infty} \rho_j = n^{-1} \sum_{-\infty}^{\infty} \gamma_j$$

that is,

$$\sigma^2(\overline{w}) = n^{-1}\gamma(1) \tag{6.3.17}$$

where $\gamma(B)$ is the autocovariance generating function defined in (3.1.10) and $\gamma(1)$ is its value when $B = B^{-1} = 1$ is substituted.

For illustration, consider the process of order $(1, d, 0)$:

$$(1 - \phi B)(w_t - \mu_w) = a_t$$

with $w_t = \nabla^d z_t$. From (3.1.11), we obtain

$$\gamma(B) = \frac{\sigma_a^2}{(1 - \phi B)(1 - \phi F)}$$

so

$$\sigma^2(\overline{w}) = n^{-1}(1 - \phi)^{-2}\sigma_a^2$$

But $\sigma_a^2 = \sigma_w^2(1 - \phi^2)$, so

$$\sigma^2(\overline{w}) = \frac{\sigma_w^2}{n}\frac{1 - \phi^2}{(1 - \phi)^2} = \frac{\sigma_w^2}{n}\frac{1 + \phi}{1 - \phi}$$

and

$$\sigma(\overline{w}) = \sigma_w \left[\frac{1 + \phi}{n(1 - \phi)}\right]^{1/2}$$

Now ϕ and σ_w^2 are estimated by r_1 and c_0, respectively, as defined in (2.1.11) and (2.1.12). Thus, for a $(1, d, 0)$ process, the required standard error is given by

$$\hat{\sigma}(\overline{w}) = \left[\frac{c_0(1 + r_1)}{n(1 - r_1)}\right]^{1/2}$$

Proceeding in this way, the expressions for $\sigma(\overline{w})$ given in Table 6.4 may be obtained.

Tentative Identification of Models A–F. Table 6.5 summarizes the models tentatively identified for Series A to F, with the preliminary parameter estimates inserted. These parameter values are used as initial guesses for the more efficient estimation methods to be described in Chapter 7.

6.3.7 Choice Between Stationary and Nonstationary Models in Doubtful Cases

As the results in Tables 6.2 and 6.5 suggest, the preliminary identification of the need for differencing and of the degree of differencing is not always easily determined. The apparent ambiguity in identifying models for Series A, C, and D (particularly with regard

TABLE 6.4 Approximate Standard Error for $\overline{w}$, where $w_t = \nabla^d z_t$ and z_t is an ARIMA Process of Order (p, d, q)

$(1, d, 0)$	$(0, d, 1)$
$\left[\frac{c_0(1+r_1)}{n(1-r_1)}\right]^{1/2}$	$\left[\frac{c_0(1+2r_1)}{n}\right]^{1/2}$
$(2, d, 0)$	$(0, d, 2)$
$\left[\frac{c_0(1+r_1)(1-2r_1^2+r_2)}{n(1-r_1)(1-r_2)}\right]^{1/2}$	$\left[\frac{c_0(1+2r_1+2r_2)}{n}\right]^{1/2}$
$(1, d, 1)$	
$\left[\frac{c_0}{n}\left(1+\frac{2r_1^2}{r_1-r_2}\right)\right]^{1/2}$	

to the degree of differencing) is, of course, more apparent than real. It arises whenever the roots of $\phi(B) = 0$ approach unity. When this happens, it becomes less and less important whether a root near unity is included in $\phi(B)$ or an additional difference is included corresponding to a unit root. A more precise evaluation is possible using the estimation procedures discussed in Chapter 7 and, in particular, the more formal unit root testing procedures to be discussed in Chapter 10. However, the following should be kept in mind:

1. From time series that are necessarily of finite length, it is never possible to *prove* that a zero of the autoregressive operator is exactly equal to unity.
2. There is, of course, no sudden transition from stationary behavior to nonstationary behavior. This can be understood by considering the behavior of the simple mixed

TABLE 6.5 Summary of Models Identified for Series A–F, with Initial Estimates Inserted

Series	Differencing	$\overline{w} \pm \hat{\sigma}(\overline{w})^a$	$\hat{\sigma}_w^2 = c_0$	Identified Model	$\hat{\sigma}_a^2$
A	Either 0	17.06 ± 0.10	0.1586	$z_t - 0.87z_{t-1} = 2.45 + a_t - 0.48a_{t-1}$	0.098
	or 1	0.002 ± 0.011	0.1364	$\nabla z_t = a_t - 0.53a_{t-1}$	0.107
B	1	-0.28 ± 0.41	52.54	$\nabla z_t = a_t + 0.09a_{t-1}$	52.2
C	Either 1	-0.035 ± 0.047	0.0532	$\nabla z_t - 0.81\nabla z_{t-1} = a_t$	0.019
	or 2	-0.003 ± 0.008	0.0198	$\nabla^2 z_t = a_t - 0.09a_{t-1} - 0.07a_{t-2}$	0.020
D	Either 0	9.13 ± 0.04	0.3620	$z_t - 0.86z_{t-1} = 1.32 + a_t$	0.093
	or 1	0.004 ± 0.017	0.0965	$\nabla z_t = a_t - 0.05a_{t-1}$	0.096
E	Either 0	46.9 ± 5.4	1382.2	$z_t - 1.32z_{t-1} + 0.63z_{t-2} = 14.9 + a_t$	289.0
	or 0	46.9 ± 5.4	1382.2	$z_t - 1.37z_{t-1} + 0.74z_{t-2} - 0.08z_{t-3} = 13.7 + a_t$	287.0
F	0	51.1 ± 1.1	139.80	$z_t + 0.32z_{t-1} - 0.18z_{t-2} = 58.3 + a_t$	115.0

[a] When $d = 0$, read z for w.

model

$$(1 - \phi_1 B)(z_t - \mu) = (1 - \theta_1 B)a_t$$

Series generated by such a model behave in a more nonstationary manner as ϕ_1 increases toward unity. For example, a series with $\phi_1 = 0.99$ can wander away from its mean μ and not return for very long periods. It is as if the attraction that the mean exerts in the series becomes less and less as ϕ_1 approaches unity, and finally, when ϕ_1 is equal to unity, the behavior of the series is completely independent of μ.

In doubtful cases, there may be an advantage in employing the nonstationary model rather than the stationary alternative (e.g., in treating a ϕ_1, whose estimate is close to unity, as being *equal* to unity). This is particularly true in forecasting and control problems. Where ϕ_1 is close to unity, we do not really know whether the mean of the series has meaning or not. Therefore, it may be advantageous to employ the nonstationary model, which does not include a fixed mean μ. If we use such a model, forecasts of future behavior will not in any way depend on an estimated mean, calculated from a previous period, which may have no relevance to the future level of the series.

6.4 MODEL MULTIPLICITY

6.4.1 Multiplicity of Autoregressive–Moving Average Models

With the normal distribution assumption, knowledge of the first and second moments of a probability distribution implies complete knowledge of the distribution. In particular, knowledge of the mean of $w_t = \nabla^d z_t$ and of its autocovariance function uniquely determines the probability structure or w_t. We now show that although this unique probability structure can be represented by a *multiplicity* of linear ARMA models, uniqueness is achieved in the model when we introduce the appropriate stationarity and invertibility restrictions.

Suppose that w_t, having autocovariance generating function $\gamma(B)$, is represented by the linear ARMA model

$$\phi(B)w_t = \theta(B)a_t \tag{6.4.1}$$

where the zeros of $\phi(B)$ and of $\theta(B)$ lie outside the unit circle. Then, this model may also be written as

$$\prod_{i=1}^{p}(1 - G_i B)w_t = \prod_{j=1}^{q}(1 - H_j B)a_t \tag{6.4.2}$$

where the G^{-1} are the roots of $\phi(B) = 0$ and H_j^{-1} are the roots of $\theta(B) = 0$, and G_i, H_j lie inside the unit circle. Using (3.1.11), the autocovariance generating function for w is

$$\gamma(B) = \prod_{i=1}^{p}(1 - G_i B)^{-1}(1 - G_i F)^{-1}\prod_{j=1}^{q}(1 - H_j B)(1 - H_j F)\sigma_a^2$$

Multiple Choice of Moving Average Parameters. Since

$$(1 - H_j B)(1 - H_j F) = H_j^2(1 - H_j^{-1} B)(1 - H_j^{-1} F)$$

it follows that any one of the stochastic models

$$\prod_{i=1}^{p}(1 - G_i B)w_t = \prod_{j=1}^{q}(1 - H_j^{\pm 1} B)ka_t$$

can have the same autocovariance generating function if the constant k is chosen appropriately. In the above, it is understood that for complex roots, reciprocals of both members of the conjugate pair will be taken (so as to always obtain *real-valued* coefficients in the MA operator). However, if a real root H is inside the unit circle, H^{-1} will lie outside, or if a complex pair, say H_1 and H_2, are inside, then the pair H_1^{-1} and H_2^{-1} will lie outside. It follows that there will be only *one stationary invertible* model of the form (6.4.2), which has a given autocovariance function.

Backward Representations. Now $\gamma(B)$ also remains unchanged if in (6.4.2) we replace $1 - G_i B$ by $1 - G_i F$ or $1 - H_j B$ by $1 - H_j F$. Thus, all the stochastic models

$$\prod_{i=1}^{p}(1 - G_i B^{\pm 1})w_t = \prod_{j=1}^{q}(1 - H_j B^{\pm 1})a_t$$

have identical autocovariance structure. However, representations containing the operator $B^{-1} = F$ refer to future w's and/or future a's, so that although stationary and invertible representations exist in which w_t is expanded in terms of future w's and a's, only one such representation, (6.4.2), exists that relates w_t entirely to *past* history.

A model form that, somewhat surprisingly, is of practical interest is that in which *all* B's are replaced by F's in (6.4.1), so that

$$\phi(F)w_t = \theta(F)e_t$$

where e_t is a sequence of independently distributed random variables having mean zero and variance $\sigma_e^2 = \sigma_a^2$. This then is a stationary invertible representation in which w_t is expressed *entirely* in terms of future w's and e's. We refer to it as the *backward* form of the process, or more simply as the *backward process*.

Equation (6.4.2) is not the most general form of a stationary invertible linear ARMA model having the autocovariance generating function $\gamma(B)$. For example, the model (6.4.2) may be multiplied on both sides by any factor $1 - QB$. Thus, the process

$$(1 - QB)\prod_{i=1}^{p}(1 - G_i B)w_t = (1 - QB)\prod_{j=1}^{q}(1 - H_j B)a_t$$

has the same autocovariance structure as (6.4.2). This fact will present no particular difficulty at the identification stage, since we will be naturally led to choose the simplest representation, and so for uniqueness we require that there be *no common factors* between the AR and MA operators in the model. However, as discussed in Chapter 7, we need to be alert to the possibility of common factors in the estimated AR and MA operators when fitting the process.

Finally, we reach the conclusion that a stationary-invertible model, in which a current value w_t is expressed only in terms of *previous* history and which contains *no common factors* between the AR and MA operators, is uniquely determined by the autocovariance structure.

Proper understanding of model multiplicity is of importance for a number of reasons:

1. We are reassured by the foregoing argument that the autocovariance function can logically be used to identify a linear stationary-invertible ARMA model that expresses w_t in terms of previous history.
2. The nature of the multiple solutions for moving average parameters obtained by equating moments is clarified.
3. The backward process

$$\phi(F)w_t = \theta(F)e_t$$

obtained by replacing B by F in the linear ARMA model, is useful in estimating values of the series that have occurred before the first observation was made.

Now we consider reasons 2 and 3 in greater detail.

6.4.2 Multiple Moment Solutions for Moving Average Parameters

In estimating the q parameters $\theta_1, \theta_2, \ldots, \theta_q$ in the MA model by equating autocovariances, we have seen that multiple solutions are obtained. To each combination of roots, there will be a corresponding linear representation, but to only one such combination will there be an invertible representation in terms of past history.

For example, consider the MA(1) process in w_t:

$$w_t = (1 - \theta_1 B)a_t$$

and suppose that $\gamma_0(w)$ and $\gamma_1(w)$ are known and we want to deduce the values of θ_1 and σ_a^2. Since

$$\gamma_0 = (1 + \theta_1^2)\sigma_a^2 \qquad \gamma_1 = -\theta_1\sigma_a^2 \qquad \gamma_k = 0 \qquad k > 1 \tag{6.4.3}$$

then

$$-\frac{\gamma_0}{\gamma_1} = \theta_1^{-1} + \theta_1$$

and if $(\theta_1 = \theta, \sigma_a^2 = \sigma^2)$ is a solution for given γ_0 and γ_1, so is $(\theta_1 = \theta^{-1}, \sigma_a^2 = \theta^2\sigma^2)$. Apparently, then, for given values of γ_0 and γ_1, there are a *pair* of possible models:

$$w_t = (1 - \theta B)a_t$$

and

$$w_t = (1 - \theta^{-1}B)\alpha_t \tag{6.4.4}$$

with $\text{var}[a_t] = \sigma_a^2$ and $\text{var}[\alpha_t] = \sigma_\alpha^2 = \sigma_a^2\theta^2$. If $-1 < \theta < 1$, then (6.4.4) is not an invertible representation. However, this model may be written as

$$w_t = [(1 - \theta^{-1}B)(-\theta F)](-\theta^{-1}B\alpha_t)$$

Thus, after setting $e_t = -\alpha_{t-1}/\theta$, the model becomes

$$w_t = (1 - \theta F)e_t \tag{6.4.5}$$

where e_t has the same variance as a_t. Thus, (6.4.5) is simply the ''backward'' process, which is dual to the forward process:

$$w_t = (1 - \theta B)a_t \tag{6.4.6}$$

Just as the shock a_t in (6.4.6) is expressible as a convergent sum of current and *previous* values of w,

$$a_t = w_t + \theta w_{t-1} + \theta^2 w_{t-2} + \cdots$$

the shock e_t in (6.4.5) is expressible as a convergent sum of current and future values of w:

$$e_t = w_t + \theta w_{t+1} + \theta^2 w_{t+2} + \cdots$$

Thus, the root θ^{-1} *would* produce an ''invertible'' process, but only if a representation of the shock e_t in terms of future values of w were permissible. The invertibility regions shown in Table 6.1 delimit acceptable values of the parameters, *given* that we express the shock in terms of *previous* history.

6.4.3 Use of the Backward Process to Determine Starting Values

Suppose that a time series $w_1, w_2, \ldots, w_n$ is available from a process

$$\phi(B)w_t = \theta(B)a_t \tag{6.4.7}$$

In Chapter 7, problems arise where we need to estimate the values w_0, w_{-1}, w_{-2}, and so on, of the series that occurred *before* the first observation was made. This happens because ''starting values'' are needed for certain basic recursive calculations used for estimating the parameters in the model. Now, suppose that we require to estimate w_{-l}, given $w_1, \ldots, w_n$. The discussion of Section 6.4.1 shows that the probability structure of $w_1, \ldots, w_n$ is equally explained by the forward model (6.4.7), or by the backward model

$$\phi(F)w_t = \theta(F)e_t \tag{6.4.8}$$

The value w_{-l}, thus, bears exactly the same probability relationship to the sequence $w_1, w_2, \ldots, w_n$, as does the value w_{n+l+1} to the sequence $w_n, w_{n-1}, w_{n-2}, \ldots, w_1$. Thus, to estimate a value $l + 1$ periods before observations started, we can first consider what would be the optimal estimate or forecast $l + 1$ periods after the series

ended, and then apply this procedure to the *reversed* series. In other words, we ''forecast'' the reversed series. We call this ''back forecasting.''

APPENDIX A6.1 EXPECTED BEHAVIOR OF THE ESTIMATED AUTOCORRELATION FUNCTION FOR A NONSTATIONARY PROCESS

Suppose that a series of N observations $z_1, z_2, \ldots, z_N$ is generated by a nonstationary $(0, 1, 1)$ process

$$\nabla z_t = (1 - \theta B)a_t$$

and the estimated autocorrelations r_k are computed, where

$$r_k = \frac{c_k}{c_0} = \frac{\sum_{t=1}^{N-k}(z_t - \bar{z})(z_{t+k} - \bar{z})}{\sum_{t=1}^{N}(z_t - \bar{z})^2}$$

Some idea of the behavior of these estimated autocorrelations may be obtained by deriving expected values for the numerator and denominator of this expression and considering the ratio. We will write, following Wichern (1973),

$$\begin{aligned}\mathcal{E}[r_k] &= \frac{E[c_k]}{E[c_0]} \\ &= \frac{\sum_{t=1}^{N-k} E[(z_t - \bar{z})(z_{t+k} - \bar{z}]}{\sum_{t=1}^{N} E[(z_t - \bar{z})^2]}\end{aligned}$$

After straightforward but tedious algebra, we find that

$$\mathcal{E}[r_k] = \frac{(N-k)[(1-\theta)^2(N^2 - 1 + 2k^2 - 4kN) - 6\theta]}{N(N-1)[(N+1)(1-\theta)^2 + 6\theta]} \tag{A6.1.1}$$

For θ close to zero, $\mathcal{E}[r_k]$ will be close to unity, but for large values of θ, it can be considerably smaller than unity, even for small values of k. Figure A6.1 illustrates this fact by showing values of $\mathcal{E}[r_k]$ for $\theta = 0.8$ with $N = 100$ and $N = 200$. Although, as anticipated for a nonstationary process, the ratios $\mathcal{E}[r_k]$ of expected values fail to damp out quickly, it will be seen that they do not approach the value 1 even for small lags.

Similar effects may be demonstrated whenever the parameters approach values where cancellation on both sides of the model would produce a stationary process. For instance, in the example above we can write the model as

$$(1 - B)z_t = [(1 - B) + \delta B]a_t$$

where $\delta = 0.2$. As δ tends to zero, the behavior of the process would be expected to come closer and closer to that of the white noise process $z_t = a_t$, for which the autocorrelation function is zero for lags $k > 0$.

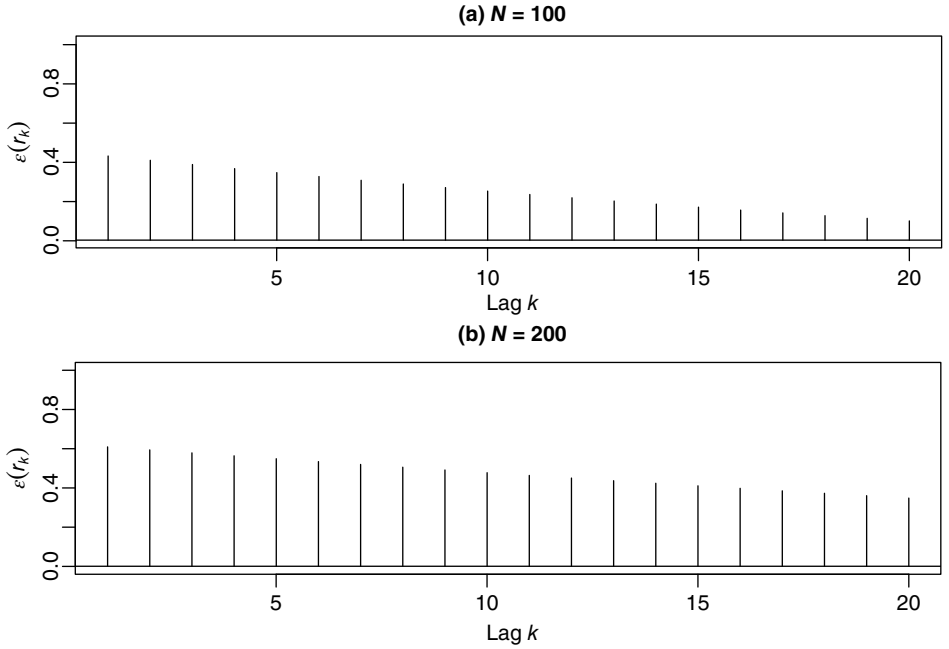

FIGURE A6.1 $\mathcal{E}[r_k] = E[c_k]/E[c_0]$ for series generated by $\nabla z_t = (1 - 0.8B)a_t$.

EXERCISES

6.1. Given the five identified models and the corresponding values of the estimated autocorrelations of $w_t = \nabla^d z_t$ in the following table:

	Identified Model			
	p	d	q	Estimated Autocorrelations
(1)	1	1	0	$r_1 = 0.72$
(2)	0	1	1	$r_1 = -0.41$
(3)	1	0	1	$r_1 = 0.40, r_2 = 0.32$
(4)	0	2	2	$r_1 = 0.62, r_2 = 0.13$
(5)	2	1	0	$r_1 = 0.93, r_2 = 0.81$

(a) Obtain preliminary estimates of the parameters analytically.

(b) Check these estimates using the charts and tables in Part Five of the book.

(c) Write down the identified models in backward shift operator notation with the preliminary estimates inserted.

6.2. For the (2, 1, 0) process considered on line (5) of Exercise 6.1, the sample mean and variance of $w_t = \nabla z_t$ are $\overline{w} = 0.23$ and $s_w^2 = 0.25$. If the series contains $N = 101$ observations,

(a) show that a constant term needs to be included in the model,

(b) express the model in the form $w_t - \phi_1 w_{t-1} - \phi_2 w_{t-2} = \theta_0 + a_t$ with numerical values inserted for the parameters, including an estimate of σ_a^2.

6.3. Consider the chemical process temperature readings referred to as Series C in this book.

(a) Plot the original series and the series of first differences using R.

(b) Use the R package to calculate and plot the ACF and PACF of this series. Repeat the calculation for the first and second differences of the series.

(c) Specify a suitable model, or models, for this series. Use the method of moments to obtain preliminary parameter estimates for the series.

6.4. Quarterly measurements of the gross domestic product (GDP) in the United Kingdom over the period 1955–1969 are included in Series P in Part Five of this book.

(a) Calculate and plot the ACF and PACF of this series.

(b) Repeat the analysis in part (a) for the first differences of the series.

(c) Identify a model for the series. Would a log transformation of the data be helpful?

(d) Obtain preliminary estimates for the parameters and for their standard errors.

(e) Obtain preliminary estimates for μ_z and σ_a^2.

6.5. Quarterly UK unemployment rate (in thousands) is part of Series P analyzed in Exercise 6.4. Repeat parts (a) to (e) of Exercise 6.4 for this series.

6.6. A time series defined by $z_t = 1000 \log_{10}(H_t)$, where H_t is the price of hogs recorded annually by the U.S. Census of Agriculture on January 1 for each of the 82 years, from 1867 to 1948 is listed as Series Q in the Collection of Time Series in Part Five. This is a well-known time series analyzed by Quenouille (1957), and others.

(a) Plot the series. Compute and plot the ACF and PACF of the series.

(b) Identify a time series model for the series.

6.7. Measurements of the annual flow of the river Nile at Ashwan from 1871 to 1970 are available as series ''Nile'' in the datasets package in R; type help(Nile) for details.

(a) Plot the series and compute the ACF and PACF for the series.

(b) Repeat the analysis in part (a) for the differenced series.

(c) Identify a model for the series. Are there any unusual features worth noting.

6.8. The file ''EuStockMarkets'' in the R datasets package contains the daily closing prices of four major European stock indices: Germany DAX (Ibis), Switzerland SMI, France CAC, and UK FTSE. The data are sampled in business time, so weekends and holidays are omitted.

(a) Plot each of the four series and compute the ACF and PACF for the series.

(b) Repeat the analysis in part (a) for the differenced series.

(c) Identify a model for the series. Are there any unusual features worth noting.

6.9. Download a time series of your choice from the Internet. Plot the time series and identify a suitable model for the series.

7

PARAMETER ESTIMATION

This chapter deals with the estimation of the parameters in ARIMA models and provides a general account of likelihood and Bayesian methods for parameter estimation. It is assumed that a suitable model of this form has been selected using the model specification tools described in Chapter 6. After the parameters have been estimated, the fitted model will be subjected to diagnostic checks and goodness-of-fit tests to be described in the next chapter. As pointed out by R. A. Fisher, for tests of goodness of fit to be relevant, it is necessary that efficient use of data should have been made in the fitting process. If this is not so, inadequacy of fit may simply arise because of the inefficient fitting and not because the form of the model is inadequate. This chapter examines in detail maximum likelihood estimation under the normality assumption and describes least-squares approximations that are suitable for many series.

It is assumed that the reader is familiar with certain basic ideas in estimation theory. Appendices A7.1 and A7.2 summarize some important results in normal distribution theory and linear least-squares that are useful for this chapter. Throughout the chapter, bold type is used to denote vectors and matrices. Thus, $\mathbf{X} = \{x_{ij}\}$ is a matrix with x_{ij} an element in the ith row and jth column, and $\mathbf{X}'$ is the transpose of the matrix $\mathbf{X}$.

7.1 STUDY OF THE LIKELIHOOD AND SUM-OF-SQUARES FUNCTIONS

7.1.1 Likelihood Function

Suppose that we have a sample of N observations, $\mathbf{z}$ with which we associate an N-dimensional random variable, whose known probability distribution $p(\mathbf{z}|\boldsymbol{\xi})$ depends on some unknown parameters $\boldsymbol{\xi}$. We use the vector $\boldsymbol{\xi}$ to denote a general set of parameters and, in particular, it could refer to the $p + q + 1$ parameters $(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2)$ of the ARIMA model.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

Before the data are available, $p(\mathbf{z}|\boldsymbol{\xi})$ will associate a density with each different outcome $\mathbf{z}$ of the experiment, for fixed $\boldsymbol{\xi}$. After the data have become available, we are led to contemplate the various values of $\boldsymbol{\xi}$ that might have given rise to the fixed set of observations $\mathbf{z}$ actually obtained. The appropriate function for this purpose is the *likelihood function* $L(\boldsymbol{\xi}|\mathbf{z})$, which is of the same form as $p(\mathbf{z}|\boldsymbol{\xi})$, but in which $\mathbf{z}$ is now fixed but $\boldsymbol{\xi}$ is variable. It is only the relative value of $L(\boldsymbol{\xi}|\mathbf{z})$ that is of interest, so that the likelihood function is usually regarded as containing an *arbitrary multiplicative constant.*

It is often convenient to work with the log-likelihood function $\ln[L(\boldsymbol{\xi}|\mathbf{z})] = l(\boldsymbol{\xi}|\mathbf{z})$, which contains an *arbitrary additive constant*. One reason that the likelihood function is of fundamental importance in estimation theory is because of the *likelihood principle*, urged on somewhat different grounds by Fisher (1956), Barnard (1949), and Birnbaum (1962). This principle says that, given that the assumed model is correct, all that the *data* have to tell us about the parameters is contained in the likelihood function, all other aspects of the data being irrelevant. From a Bayesian point of view, the likelihood function is equally important, since it is the component in the posterior distribution of the parameters that comes from the data.

For a complete understanding of the parameter estimation in a specific case, it is necessary to carefully study of the likelihood function, or in the Bayesian framework, the posterior distribution of the parameters, which in the cases we consider, is dominated by the likelihood. In many examples, for moderate and large samples, the log-likelihood function will be unimodal and can be approximated adequately over a sufficiently extensive region near the maximum by a quadratic function. In such cases, the log-likelihood function can be described by its maximum and its second derivatives at the maximum. The values of the parameters that maximize the likelihood function, or equivalently the log-likelihood function, are called *maximum likelihood (ML) estimates*. The second derivatives of the log-likelihood function provide measures of ''spread'' of the likelihood function and can be used to calculate approximate standard errors for the estimates.

The limiting properties of maximum likelihood estimates are usually established for independent observations. But as was shown by Whittle (1953), they may be extended to cover stationary time series. Other early literature on the parameter estimation in time series models includes Barnard et al. (1962), Bartlett (1955), Durbin (1960), Grenander and Rosenblatt (1957), Hannan (1960), and Quenouille (1942, 1957).

7.1.2 Conditional Likelihood for an ARIMA Process

Let us suppose that the $N = n + d$ original observations $\mathbf{z}$ form a time series that we denote by $z_{-d+1}, \ldots, z_0, z_1, z_2, \ldots, z_n$. We assume that this series is generated by an ARIMA(p, d, q) model. From these observations, we can generate a series $\mathbf{w}$ of $n = N - d$ differences $w_1, w_2, \ldots, w_n$ where $w_t = \nabla^d z_t$. Thus, the general problem of fitting the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ of the ARIMA model (6.1.1) is equivalent to fitting to the w_t's, the stationary and invertible[1] ARMA(p, q) model, which may be written as

$$a_t = \tilde{w}_t - \phi_1\tilde{w}_{t-1} - \phi_2\tilde{w}_{t-2} - \cdots - \phi_p\tilde{w}_{t-p} + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q} \tag{7.1.1}$$

where $\tilde{w}_t = w_t - \mu$ are the mean-centered observations.

[1] Special care is needed to ensure that estimate lies in the invertible region. See Appendix A7.7.

For $d > 0$, it is often appropriate to assume that $\mu = 0$. When this is not appropriate, we assume that the series mean $\bar{w} = \sum_{t=1}^{n} w_t/n$ is substituted for μ. For many sample sizes common in practice, this approximation will be adequate. However, if desired, μ can be included as an additional parameter to be estimated.

The a_t's cannot be calculated immediately from (7.1.1) because of the difficulty of starting up the difference equation. However, suppose that the p values $\mathbf{w}_*$ of the w_t's and the q values $\mathbf{a}_*$ of the a_t's prior to the start of the w_t series were given. Then, for any choice of parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$, we could calculate successively a set of values $a_t(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{w}_*, \mathbf{a}_*, \mathbf{w})$, $t = 1, 2, \ldots, n$. Now, assuming that the a_t's are normally distributed, their probability density is

$$p(a_1, a_2, \ldots, a_n) \propto (\sigma_a^2)^{-n/2} \exp\left[-\left(\sum_{t=1}^{n} \frac{a_t^2}{2\sigma_a^2}\right)\right]$$

Given the data $\mathbf{w}$, the log-likelihood associated with the parameter values $(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2)$, *conditional* on the choice of $(\mathbf{w}_*, \mathbf{a}_*)$, would then be

$$l_*(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = -\frac{n}{2}\ln(\sigma_a^2) - \frac{S_*(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_a^2} \tag{7.1.2}$$

where

$$S_*(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} a_t^2(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{w}_*, \mathbf{a}_*, \mathbf{w}) \tag{7.1.3}$$

In the above equations, a subscript asterisk is used on the likelihood and sum-of-squares functions to emphasize that they are conditional on the choice of the starting values. We notice that the conditional log-likelihood l_* involves the data only through the conditional *sum-of-squares function*. It follows that contours of l_* for any fixed value of σ_a^2 in the space of $(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2)$ are contours of S_*, that these maximum likelihood estimates are the same as the least-squares estimates, and that in general, we can, on the normal assumption, study the behavior of the conditional likelihood by studying the conditional sum-of-squares function. In particular for any fixed σ_a^2, l_* is a linear function of S_*. The parameter values obtained by minimizing the conditional sum-of-squares function $S_*(\boldsymbol{\phi}, \boldsymbol{\theta})$ will be called *conditional least-squares estimates.*

7.1.3 Choice of Starting Values for Conditional Calculation

We will shortly discuss the calculation of the unconditional likelihood, which, strictly, is what we need for parameter estimation. However, when n is moderate or large, a sufficient approximation to the unconditional likelihood is often obtained by using the conditional likelihood with suitable values substituted for the elements of $\mathbf{w}_*$ and $\mathbf{a}_*$ in (7.1.3). One procedure is to set the elements of $\mathbf{w}_*$ and of $\mathbf{a}_*$ equal to their unconditional expectations. The unconditional expectations of the elements of $\mathbf{a}_*$ are zero, and if the model contains no deterministic part, and in particular if $\mu = 0$, the unconditional expectations of the elements

TABLE 7.1 Sum-of-Squares Functions for the Model $\nabla z_t = (1 - \theta B)a_t$ Fitted to the IBM Data

θ	$\lambda = 1 - \theta$	$S_*(\theta)$	$S(\theta)$	θ	$\lambda = 1 - \theta$	$S_*(\theta)$	$S(\theta)$
−0.5	1.5	23,929	23,928	0.1	0.9	19,896	19,896
−0.4	1.4	21,595	21,595	0.2	0.8	20,851	20,851
−0.3	1.3	20,222	20,222	0.3	0.7	22,315	22,314
−0.2	1.2	19,483	19,483	0.4	0.6	24,471	24,468
−0.1	1.1	19,220	19,220	0.5	0.5	27,694	27,691
0.0	1.0	19,363	19,363				

of $\mathbf{w}_*$ will also be zero[2]. However, this approximation can be poor if some of the roots of $\phi(B) = 0$ are close to the boundary of the unit circle, so that the process approaches nonstationarity. This is also true if some of the roots of $\theta(B) = 0$ are close to the boundary of the invertibility region. Setting the presample values equal to zero could in these cases introduce a large transient, which is slow to die out. For a pure AR(p) model, a more reliable approximation procedure, and one we employ sometimes, is to use (7.1.1) to calculate the a_t's *from* a_{p+1} *onward*, thus using actual values of the w_t's throughout. Using this method, we have only $n - p = N - p - d$ values of a_t, but the slight loss of information will be unimportant for long series.

For seasonal series, discussed in Chapter 9, the conditional approximation is not always satisfactory and the unconditional calculation becomes necessary. Inclusion of the determinant in the unconditional likelihood function can also be important for seasonal time series.

Example: IMA(0, 1, 1) Process. To illustrate the recursive calculation of the conditional sum of squares S_*, we consider the IMA(0, 1, 1) model tentatively identified in Section 6.4 for the IBM data in Series B. The model is

$$\nabla z_t = (1 - \theta B)a_t \qquad -1 < \theta < 1 \tag{7.1.4}$$

so that $a_t = w_t + \theta a_{t-1}$, where $w_t = \nabla z_t$ and $E[w_t] = 0$. Thus, for the particular parameter value $\theta = 0.5$, the a_t's are calculated recursively from

$$a_t = w_t + 0.5a_{t-1}$$

setting the initial value a_0 equal to zero. Proceeding in this way, we find that

$$S_*(0.5) = \sum_{t=1}^{368} a_t^2(\theta = 0.5 | a_0 = 0) = 27,694$$

The conditional sums of squares $S_*(\theta)$ are shown in Table 7.1 for values of θ from −0.5 to +0.5 in steps of 0.1. We note that $S_*(\theta)$ has its minimum for $\theta = -0.1$. This is consistent with the preliminary moment estimate of −0.09 derived for this series in Chapter 6.

[2]If the assumption $E|w_t| = \mu \neq 0$ is appropriate, we can substitute $\bar{w}$ for each of the elements of $\mathbf{w}_*$.

7.1.4 Unconditional Likelihood, Sum-of-Squares Function, and Least-Squares Estimates

Assuming that the $N = n + d$ observations are generated by an ARIMA model, the unconditional log-likelihood is given by

$$l(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = f(\boldsymbol{\phi}, \boldsymbol{\theta}) - \frac{n}{2}\ln(\sigma_a^2) - \frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_a^2} \tag{7.1.5}$$

where $f(\boldsymbol{\phi}, \boldsymbol{\theta})$ involves the determinant in the joint density of the w_t's and is a function of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. The *unconditional sum-of-squares function* is given by

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} [a_t|\mathbf{w}, \boldsymbol{\phi}, \boldsymbol{\theta}]^2 + [\mathbf{e}_*]'\boldsymbol{\Omega}^{-1}[\mathbf{e}_*] \tag{7.1.6}$$

where $[a_t|\mathbf{w}, \boldsymbol{\phi}, \boldsymbol{\theta}] = E[a_t|\mathbf{w}, \boldsymbol{\phi}, \boldsymbol{\theta}]$ denotes the expectation of a_t conditional on $\mathbf{w}$, $\boldsymbol{\phi}$, and $\boldsymbol{\theta}$. When the meaning is clear from the context, we will further abbreviate this conditional expectation to $[a_t]$. In (7.1.6),

$$\mathbf{e}_* = (\bar{w}_{1-p}, \ldots, \bar{w}_0, a_{1-q}, \ldots, a_0)'$$

represents the $p + q$ initial values of the $\bar{w}_t$ and a_t prior to $t = 1$, $\boldsymbol{\Omega}\sigma_a^2 = \text{cov}[\mathbf{e}_*]$ is the covariance matrix of $\mathbf{e}_*$, and

$$[\mathbf{e}_*] = ([\tilde{w}_{1-p}], \ldots, [\tilde{w}_0], [a_{1-q}], \ldots, [a_0])'$$

denotes the vector of conditional expectations (''back-forecasts'') of the initial values, given $\mathbf{w}$, $\boldsymbol{\phi}$, and $\boldsymbol{\theta}$. An alternative way to represent $S(\boldsymbol{\phi}, \boldsymbol{\theta})$ is as

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=-\infty}^{n} [a_t]^2$$

which in comparison with (7.1.6) indicates that $\Sigma_{t=-\infty}^{0}[a_t]^2 = [\mathbf{e}_*]'\boldsymbol{\Omega}^{-1}[\mathbf{e}_*]$.

Usually, $f(\boldsymbol{\phi}, \boldsymbol{\theta})$ is of importance only for small n. For moderate and large values of n, (7.1.5) is dominated by $S(\boldsymbol{\phi}, \boldsymbol{\theta})/2\sigma_a^2$, and thus the contours of the unconditional sum-of-squares function in the space of the parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$ are very nearly contours of the likelihood and log-likelihood. It follows, in particular, that the parameter estimates obtained by minimizing the sum of squares (7.1.6), which we call (*unconditional* or *exact*) *least-squares estimates*, will usually provide very close approximations to the maximum likelihood estimates. From a Bayesian viewpoint, on assumptions discussed in Section 7.5, for all AR(p) and MA(q), essentially the posterior density is a function only of $S(\boldsymbol{\phi}, \boldsymbol{\theta})$. Hence, very nearly the least-squares estimates are those with maximum posterior density. In the remainder of this section and in Section 7.1.5, the main emphasis will be on the unconditional sum-of-squares function $S(\boldsymbol{\phi}, \boldsymbol{\theta})$ in (7.1.6), and its use in calculating least-squares estimates. An alternate method for calculation of the unconditional sum of squares and likelihood functions based on the state-space model and innovations approach will be discussed in Section 7.4.

In the calculation of the unconditional sum of squares, the $[a_t]$'s are computed recursively by taking conditional expectations in (7.1.1). A preliminary back-calculation provides the

values $[w_{-j}]$ and $[a_{-j}]$, $j = 0, 1, 2, \ldots$ (i.e., the back-forecasts) needed to start off the forward recursion.

Calculation of the Unconditional Sum of Squares for a Moving Average Process. For illustration, we reconsider the IBM stock price example using only the first 10 values of the series.[3] For the IMA(0, 1, 1) model, the only back-forecast that is needed for $S(\theta)$ is $[a_0]$. We begin by describing an *approximate*, but nevertheless accurate, method to obtain $[a_0]$. Recall from Section 6.4.3 that the model for w_t may be written in either the forward or backward forms:

$$w_t = (1 - \theta B)a_t \qquad w_t = (1 - \theta F)e_t$$

and where again $\mu = E[w_t]$ is assumed equal to zero. Hence, we can write

$$[e_t] = [w_t] + \theta[e_{t+1}] \tag{7.1.7}$$

$$[a_t] = [w_t] + \theta[a_{t-1}] \tag{7.1.8}$$

where $[w_t] = w_t$ for $t = 1, 2, \ldots, n$ and is the back-forecast of w_t for $t \leq 0$. These are the two basic equations that we need in the computations. A convenient format for the calculations is shown in Table 7.2. We begin by entering in the table what we know:

1. The data values $z_0, z_1, \ldots, z_9$, from which we can calculate the first differences $w_1, w_2, \ldots, w_9$.
2. The values $[e_0], [e_{-1}], \ldots$, which are zero, since $e_0, e_{-1}, \ldots$ are distributed independently of **w**.
3. The values $[a_{-1}], [a_{-2}], \ldots$, which are zero, because for any MA(q) process, $a_{-q}, a_{-q-1}, \ldots$ are distributed independently of **w**. However, note that $[a_0], [a_{-1}], \ldots, [a_{-q+1}]$ will be nonzero and can be obtained by back-forecasting. Thus, in the present example, $[a_0]$ is computed this way.

Beginning at the end of the series, (7.1.7) is now used to compute the $[e_t]$'s for $t = 9, 8, 7, \ldots, 1$. We start the backward process by setting $[e_{10}] = 0$. The effect of this approximation will be to introduce a transient into the system. However, for series of moderate length, the effect will typically be negligible by the time the beginning of the series is reached and thus will not affect the calculation of the a_t's. If desired, the adequacy of this approximation can be checked in any given case by performing a second iterative cycle.

Thus, to start the recursion in Table 7.2, in the row corresponding to $t = 9$, we enter a zero in the sixth column for the unknown value $0.5[e_{10}]$. Then, using (7.1.7), we obtain

$$\begin{aligned}[e_9] &= [w_9] + 0.5[e_{10}] \\ &= w_9 + 0 = -3\end{aligned}$$

[3]In practice, of course, useful parameter estimates could not be obtained from as few as 10 observations. We utilize this data subset merely to illustrate the calculations.

TABLE 7.2 Calculation of the $[a]$'s from the First 10 Values of Series B, Using $\theta = 0.5$

t	z_t	$[a_t]$	$0.5[a_{t-1}]$	$[w_t]$	$0.5[e_{t+1}]$	$[e_t]$	u_t
−1	[458.4]	0	0	0	0	0	
0	460	1.6	0	1.6	−1.6	0	−2.1
1	457	−2.2	0.8	−3.0	−0.1	−3.1	−4.1
2	452	−6.1	−1.1	−5.0	4.8	−0.2	−2.3
3	459	3.9	−3.0	7.0	2.6	9.6	8.5
4	462	5.0	2.0	3.0	2.3	5.3	9.5
5	459	−0.5	2.5	−3.0	7.6	4.6	9.2
6	463	3.7	−0.2	4.0	11.1	15.1	19.4
7	479	17.9	1.9	16.0	6.2	22.2	31.4
8	493	22.9	9.0	14.0	−1.5	12.5	27.5
9	490	8.5	11.5	−3.0	0	−3.0	8.5

so $0.5[e_9] = -1.5$ can be entered in the line $t = 8$, which enables us to compute $[e_8]$, and so on. Finally, we obtain

$$[e_0] = [w_0] + \theta[e_1]$$

that is, $0 = [w_0] - 1.6$, which gives $[w_0] = 1.6$, and thereafter $[w_{-h}] = 0, h = 1, 2, 3, \ldots$. Now, using (7.1.8) with $t = 0$, we obtain

$$[a_0] = [w_0] + \theta[a_{-1}] = 1.6 + (0.5)(0) = 1.6$$

and we can then continue the forward calculations of the remaining $[a_t]$'s, leading to $S(0.5) = \Sigma_{t=0}^{9}[a_t|0.5, \mathbf{w}]^2 = 1016.406$.

An alternative method that yields exact estimates of the presample values is presented in Appendix A7.3. For the model considered above, this method involves first computing the values $a_t(a_0 = 0)$, which we abbreviate as a_t^0, by the conditional method as $a_t^0 = w_t + \theta a_{t-1}^0, t = 1, 2, \ldots, n$, using $a_0^0 = 0$ as the initial value. Then a backward recursion is performed to obtain $u_t = a_t^0 + \theta u_{t+1}$, beginning from $t = n$, down to $t = 0$, with $u_{n+1} = 0$ as the starting value. Finally, then, the exact estimate of $[a_0]$ is given by $[a_0] = -u_0(1 - \theta^2)/(1 - \theta^{2(n+1)})$. Using this starting value, the $[a_t]$ are computed from the forward recursion $[a_t] = w_t + \theta[a_{t-1}], t = 1, 2, \ldots, n$, as in (7.1.8) and the exact sum of squares becomes $S(\theta) = \Sigma_{t=0}^{n}[a_t]^2$.

In the above example, by first computing the a_t^0 using a forward recursion setting $a_0 = 0$, we obtain the values of u_t by the backward recursion for $t = 9, 8, \ldots, 0$, displayed in the final column of Table 7.2. Hence, we obtain the exact estimate of a_0 as $[a_0] = -u_0(1 - \theta^2)/(1 - \theta^{2(n+1)}) = 1.549$. This value is very close to the approximate value of 1.545 obtained by the backward model approach, and the small difference has essentially no effect on the calculation of the remaining values $[a_t]$. Using the exact method for the entire series, we find that the unconditional sum of squares for $\theta = 0.5$ is

$$S(0.5) = \sum_{t=0}^{368}[a_t|0.5, \mathbf{w}]^2 = 27,691$$

which for this particular example is very close to the conditional value $S_*(0.5) = 27,694$. The unconditional sum of squares $S(\theta)$, for values of θ between -0.5 and $+0.5$, have been added to Table 7.1 and are very close to the conditional values $S_*(\theta)$ computed earlier.

7.1.5 General Procedure for Calculating the Unconditional Sum of Squares

In the above example, w_t was a first-order moving average process, with zero mean. It followed that all forecasts for lead times greater than 1 were zero and consequently that only one preliminary value (the back-forecast $[w_0] = 1.6$) was required to start the recursive calculations using the approximate approach, and only one value $[a_0]$ in the exact approach. For a qth-order moving average process, q nonzero preliminary values $[w_0], [w_{-1}], \ldots, [w_{1-q}]$ would be needed, or equivalently, the q values $[a_0], [a_{-1}], \ldots, [a_{1-q}]$ in the exact approach, with $S(\boldsymbol{\theta}) = \sum_{t=1-q}^{n} [a_t]^2$. Special procedures, which we discuss in Section 7.3.1, are available for estimating parameters in autoregressive models. However, we show in Appendix A7.3 that the procedure described in this section can supply the unconditional sum of squares for any ARIMA model.

Specifically, suppose that the w_t's are generated by the stationary forward model

$$\phi(B)\tilde{w}_t = \theta(B)a_t \tag{7.1.9}$$

where $w_t = \nabla^d z_t$ and $\tilde{w}_t = w_t - \mu$. Then, they could equally well have been generated by the backward model

$$\phi(F)\tilde{w}_t = \theta(F)e_t \tag{7.1.10}$$

As before, in the approximate method that utilizes the backward model, we could first employ (7.1.10) to supply back-forecasts $[\tilde{w}_{-j}|\mathbf{w}, \boldsymbol{\phi}, \boldsymbol{\theta}]$. Theoretically, the presence of the autoregressive operator ensures a series of such estimates that is infinite in extent. However, assuming stationarity, the estimates $[\tilde{w}_t]$ at and beyond some point $t = -Q$, with Q of moderate size, become essentially equal to zero. Thus, to a sufficient approximation, we can write

$$\tilde{w}_t = \phi^{-1}(B)\theta(B)a_t = \sum_{j=0}^{\infty} \psi_j a_{t-j} \simeq \sum_{j=0}^{Q} \psi_j a_{t-j}$$

This means that the original mixed process could be replaced by a moving average process of order Q, and the procedure for moving averages outlined in Section 7.1.4 may be used.

Thus, in general, the dual set of equations for generating the conditional expectations $[a_t|\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{w}]$ is obtained by taking conditional expectations in (7.1.10) and (7.1.9). That is,

$$\phi(F)[\tilde{w}_t] = \theta(F)[e_t] \tag{7.1.11}$$

is first used to generate the backward forecasts and then

$$\phi(B)[\tilde{w}_t] = \theta(B)[a_t] \tag{7.1.12}$$

is used to generate the $[a_t]$'s. If we find that the forecasts are negligible in magnitude beyond some lead time Q, the recursive calculation goes forward with

$$\begin{aligned} [e_{-j}|\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{w}] &= 0 \qquad j = 0, 1, 2, \ldots \\ [a_{-j}|\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{w}] &= 0 \qquad j > Q - 1 \end{aligned} \tag{7.1.13}$$

and the sum of squares is approximated by $S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1-Q}^{n} [a_t]^2$. As mentioned earlier, a second iterative cycle in this approximate method could be used, if desired.

Alternatively, for the general model (7.1.9), the exact method discussed in Appendix A7.3 can be used to obtain the sum of squares as

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1-q}^{n} [a_t]^2 + ([\mathbf{w}_*] - \mathbf{C}'[\mathbf{a}_*])'\mathbf{K}^{-1}([\mathbf{w}_*] - \mathbf{C}'[\mathbf{a}_*]) \tag{7.1.14}$$

Here, the vectors $[\mathbf{w}_*]' = ([\tilde{w}_{1-p}], \ldots, [\tilde{w}_0])$ and $[\mathbf{a}_*]' = ([a_{1-q}], \ldots, [a_0])$ are the exact back-forecasted values obtained as in (A7.3.12). They are given by $[\mathbf{e}_*] = ([\mathbf{w}_*]', [\mathbf{a}_*]')' = \mathbf{D}^{-1}\mathbf{F}'\mathbf{u}$, where the values u_t, $t = 1, \ldots, n$ of the vector $\mathbf{u}$ are obtained through the backward recursion $u_t = a_t^0 + \theta_1 u_{t+1} + \cdots + \theta_q u_{t+q}$ with zero initial values $u_{n+1} = \cdots = u_{n+q} = 0$, and the a_t^0 are the conditional values of the a_t computed from (7.1.12) using zero initial values, $a_{1-q}^0 = \cdots = a_0^0 = 0$ and $\tilde{w}_{1-p}^0 = \cdots = \tilde{w}_0^0 = 0$. After solving the equations $\mathbf{D}[\mathbf{e}_*] = \mathbf{F}'\mathbf{u}$, as described in (A7.3.12), the exact $[a_t]$'s are then calculated through the recursion

$$[a_t] = [\tilde{w}_t] - \phi_1[\tilde{w}_{t-1}] - \cdots - \phi_p[\tilde{w}_{t-p}] + \theta_1[a_{t-1}] + \cdots + \theta_q[a_{t-q}] \tag{7.1.15}$$

for $t = 1, 2, \ldots, n$ using the exact back-forecasts as starting values, with $[\tilde{w}_t] = \tilde{w}_t$ for $1 \le t \le n$. The matrices $\mathbf{C}$, $\mathbf{K}$, $\mathbf{D}$, and $\mathbf{F}$ necessary for the computation in (7.1.14) are defined explicitly in Appendix A7.3.

Comment on the Approximation. We saw that for the IMA(0, 1, 1) model fitted to the IBM Series B, the *conditional* sums of squares provides a very close approximation to the unconditional value. This will generally be the case for sufficiently long nonseasonal time series. However, as is discussed further in Chapter 9, for seasonal series, in particular, the conditional approximation becomes less satisfactory and the unconditional sum of squares should ordinarily be computed. Moreover, including the determinant in the likelihood function to obtain exact maximum likelihood estimates of the parameters can be beneficial if the roots of the moving average operator are close to the unit circle.

Simulation studies have been performed by Dent and Min (1978) and Ansley and Newbold (1980) to empirically investigate and compare the performance of the conditional least-squares, unconditional least-squares, and maximum likelihood estimators for ARMA models. Generally, the conditional and unconditional least-squares estimators serve as satisfactory approximations to the maximum likelihood estimator for large-sample sizes. However, the simulation evidence suggests a preference for the maximum likelihood estimator for small- or moderate-sample sizes, especially if the moving average operator has a root close to the boundary of the invertibility region. Some additional information on the relative performance of the different estimators was provided by Hillmer and Tiao (1979) and Osborn (1982), who examined the *expected values* of the conditional sum of squares, the unconditional sum of squares, and the log-likelihood for an MA(1) model, as functions of the unknown parameter θ, for different sample sizes n. These studies provide an idea of

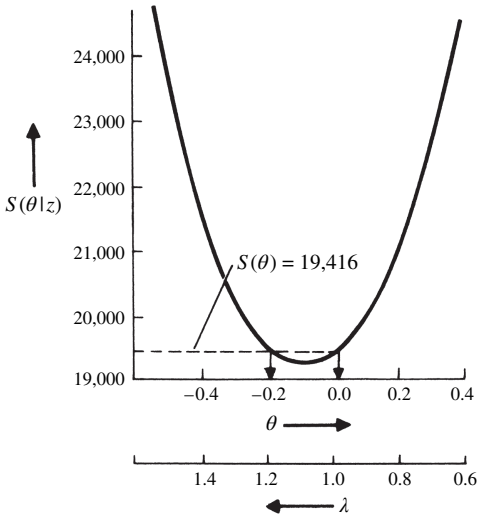

FIGURE 7.1 Plot of $S(\theta)$ for Series **B**.

how the corresponding estimators will behave for various sample sizes, and the results are consistent with those obtained from simulation studies.

7.1.6 Graphical Study of the Sum-of-Squares Function

The sum-of-squares function $S(\theta)$ for the IBM data given in Table 7.1 is plotted in Figure 7.1. The overall minimum sum of squares is at about $\theta = -0.09$ ($\lambda = 1.09$), which is the least-squares estimate and, on the assumption of normality, a close approximation to the *maximum likelihood estimate* of the parameter θ.

The graphical study of the sum-of-squares functions is readily extended to two parameters by evaluating the sum of squares over a suitable grid of parameter values and plotting contours. As discussed earlier, on the assumption of normality, the contours are very nearly likelihood contours. Figure 7.2 shows a grid of $S(\lambda_0, \lambda_1)$ values for Series B fitted with the IMA(0, 2, 2) model:

$$\begin{aligned}\nabla^2 z_t &= (1 - \theta_1 B - \theta_2 B^2)a_t \\ &= [1 - (2 - \lambda_0 - \lambda_1)B - (\lambda_0 - 1)B^2]a_t \end{aligned} \tag{7.1.16}$$

or in the form

$$\nabla^2 z_t = (\lambda_0 \nabla + \lambda_1)a_{t-1} + \nabla^2 a_t$$

The minimum sum of squares in Figure 7.2 is at about $\hat{\lambda}_0 = 1.09$ and $\hat{\lambda}_1 = 0.0$. The plot thus confirms that the preferred model in this case is an IMA(0, 1, 1) process. The device illustrated here, of fitting a model somewhat more elaborate than that expected to be needed, can provide a useful confirmation of the original identification. The elaboration of the model should be made, of course, in the direction ''feared'' to be necessary.

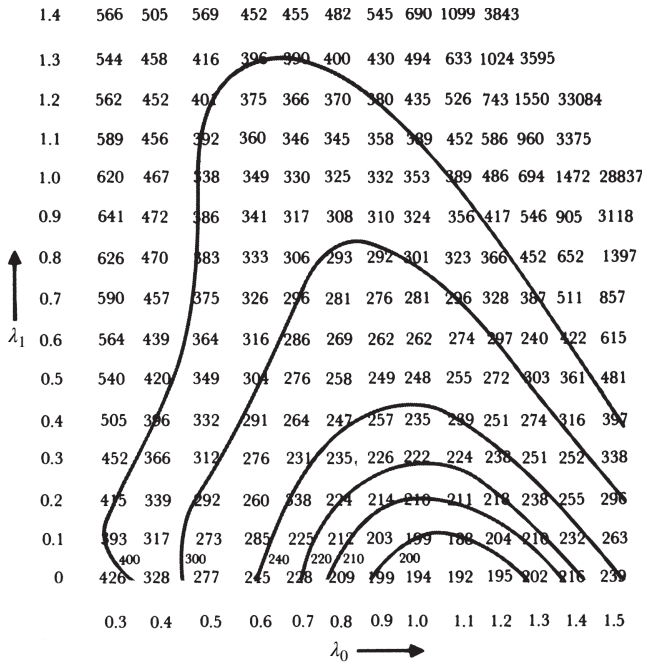

FIGURE 7.2 Values of $S(\lambda_0, \lambda_1) \times 10^{-2}$ for Series B on a grid of (λ_0, λ_1) values and approximate contours.

Three Parameters. When we wish to study models with three parameters, two-dimensional contour diagrams for a number of values of the third parameter can be drawn. For illustration, part of such a series of diagrams is shown in Figure 7.3 for Series A, C, and D. In each case, the ''elaborated'' model

$$\begin{aligned}\nabla^2 z_t &= (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3) a_t \\ &= [1 - (2 - \lambda_{-1} - \lambda_0 - \lambda_1) B - (\lambda_0 + 2\lambda_{-1} - 1) B^2 + \lambda_{-1} B^3] a_t\end{aligned}$$

or

$$\nabla^2 z_t = (\lambda_{-1} \nabla^2 + \lambda_0 \nabla + \lambda_1) a_{t-1} + \nabla^2 a_t$$

has been fitted, leading to the conclusion that the best-fitting models of this type[4] are as shown in Table 7.3.

The inclusion of additional parameters (particularly λ_{-1}) in this fitting process is not strictly necessary, but we have included them to illustrate the effect of overfitting and to show how closely our identification seems to be confirmed for these series.

[4]We show later in Section 7.2.5 that slightly better fits are obtained in some cases with closely related models containing ''stationary'' autoregressive terms.

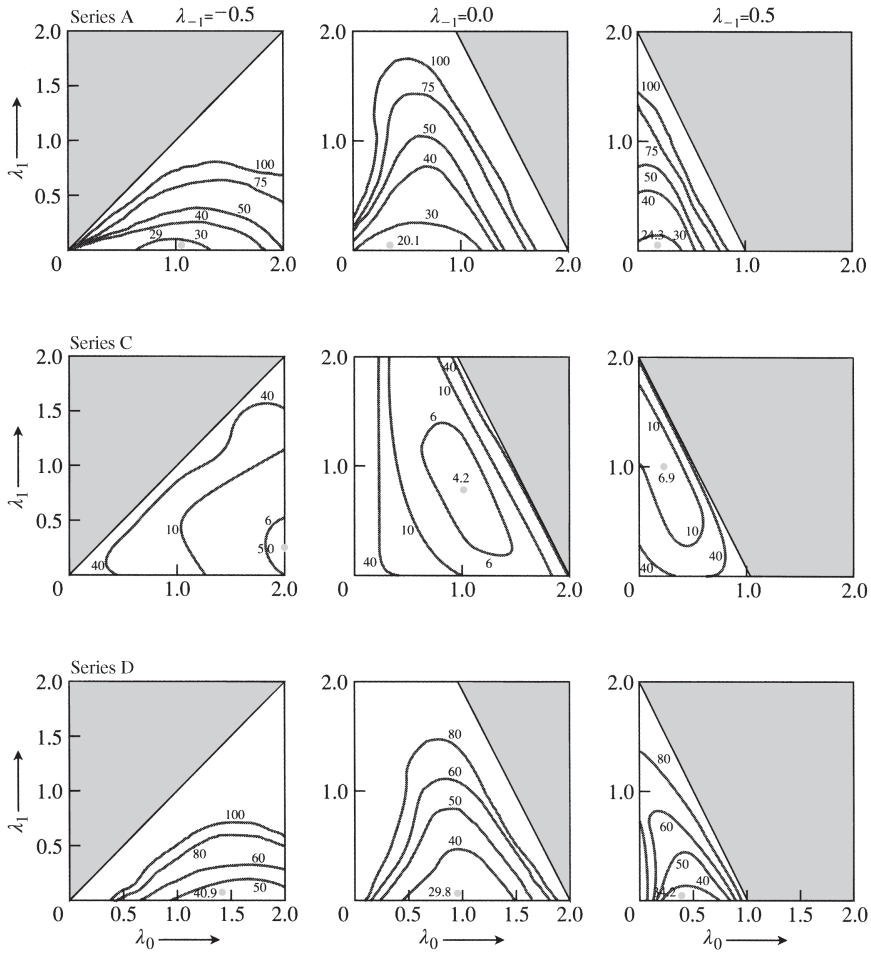

FIGURE 7.3 Sum-of-squares contours for Series A, C, and D (shaded lines indicate boundaries of the invertibility regions).

TABLE 7.3 IMA Models Fitted to Series A, C, and D

Series	$\hat{\lambda}_{-1}$	$\hat{\lambda}_0$	$\hat{\lambda}_1$	Fitted Series
A	0	0.3	0.0	$\nabla z_t = 0.3a_{t-1} + \nabla a_t$
C	0	1.1	0.8	$\nabla^2 z_t = 1.1\nabla a_{t-1} + 0.8a_{t-1} + \nabla^2 a_t$
D	0	0.9	0.0	$\nabla z_t = 0.9a_{t-1} + \nabla a_t$

7.1.7 Examination of the Likelihood Function and Confidence Regions

The likelihood function is not, of course, plotted merely to indicate maximum likelihood values. The graph of this function contains the totality of information that comes from the data. In some fields of study, cases can occur where the likelihood function has two or more peaks and also where the likelihood function contains sharp ridges and spikes. In each such

case, the likelihood function is trying to tell us something that we need to know. Thus, the existence of two peaks of approximately equal heights implies that there are two sets of parameter values that might explain the data. The existence of obliquely oriented ridges means that a value of one parameter, considerably different from its maximum likelihood value, could explain the data if accompanied by a value of the other parameter, which deviated appropriately. To understand the estimation fully, it is thus useful to examine the likelihood function both analytically and graphically.

Need for Care in Interpreting the Likelihood Function. Care is needed in interpreting the likelihood function. For example, results discussed later, which assume that the log-likelihood is approximately quadratic near its maximum, will clearly not apply to the three-parameter cases depicted in Figure 7.3. However, these examples are exceptional because here we are deliberately *overfitting* the model. If the simpler model is justified, we should *expect* to find the likelihood function contours truncated near its maximum by a boundary in the higher dimensional parameter space. However, quadratic approximations *could* be used if the simpler *identified* model rather than the overparameterized model was fitted.

Special care is needed when the maximum of the likelihood function may be on or near a boundary. Consider the situation shown in Figure 7.4 and suppose we know a priori that a parameter $\beta > \beta_0$. The maximum likelihood within the permissible range of β is at B, where $\beta = \beta_0$, not at A or at C. It will be noticed that the first derivative of the likelihood is in this case *nonzero* at the maximum likelihood value and that the quadratic approximation is certainly not an adequate representation of the likelihood.

When a class of estimation problems are examined initially, it is important to plot the likelihood function to identify potential issues. After the behavior of a potential model is well understood, and knowledge of the situation indicates that it is appropriate to do so, we may take certain shortcuts, which we now consider. We begin by considering expressions for the variances and covariances of maximum likelihood estimates, appropriate when the log-likelihood is approximately quadratic and the sample size is moderately large.

In what follows, it is convenient to define a vector $\boldsymbol{\beta}$ whose $k = p + q$ elements are the autoregressive and moving average parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. Thus, the complete set of

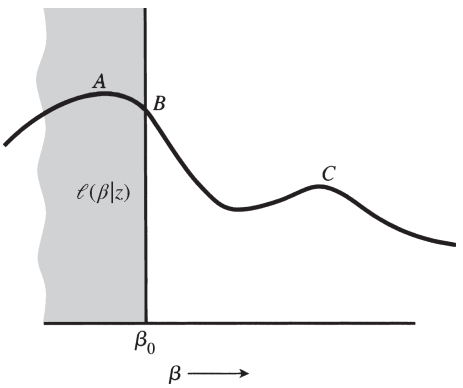

FIGURE 7.4 Hypothetical likelihood function with a constraint $\beta > \beta_0$.

$p + q + 1 = k + 1$ parameters of the ARMA process may be written as $\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2$; or as $\boldsymbol{\beta}, \sigma_a^2$; or simply as $\boldsymbol{\xi}$.

Variances and Covariances of ML Estimates. For the appropriately parameterized ARMA model, it will often happen that over the relevant[5] region of the parameter space, the log-likelihood is approximately quadratic in the elements of $\boldsymbol{\beta}$ (i.e., of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$), so that

$$l(\boldsymbol{\xi}) = l(\boldsymbol{\beta}, \sigma_a^2) \simeq l(\hat{\boldsymbol{\beta}}, \sigma_a^2) + \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} l_{ij}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j) \tag{7.1.17}$$

where, to the approximation considered, the derivatives

$$l_{ij} = \frac{\partial^2 l(\boldsymbol{\beta}, \sigma_a^2)}{\partial \beta_i \partial \beta_j} \tag{7.1.18}$$

are constant. For large n, the influence of the term $f(\boldsymbol{\phi}, \boldsymbol{\theta})$ in (7.1.5) can be ignored in most cases. Hence, $l(\boldsymbol{\beta}, \sigma_a^2)$ will be essentially quadratic in $\boldsymbol{\beta}$ if this is true for $S(\boldsymbol{\beta})$. Alternatively, $l(\boldsymbol{\beta}, \sigma_a^2)$ will be essentially quadratic in $\boldsymbol{\beta}$ if the conditional expectations $[a_t|\boldsymbol{\beta}, \mathbf{w}]$ in (7.1.6) are approximately locally linear in the elements of $\boldsymbol{\beta}$. Thus, for moderate- and large-sample sizes n, when the local quadratic approximation (7.1.17) is adequate, useful approximations to the variances and covariances of the estimates and approximate confidence regions may be obtained.

Information Matrix for the Parameters $\boldsymbol{\beta}$. The $(k \times k)$ matrix $-\{E[l_{ij}]\} = \mathbf{I}(\boldsymbol{\beta})$ is referred to (Fisher, 1956; Whittle, 1953) as the *information matrix* for the parameters $\boldsymbol{\beta}$, where the expectation is taken over the distribution of $\mathbf{w}$. For a given value of σ_a^2, the *variance–covariance* matrix $\mathbf{V}(\hat{\boldsymbol{\beta}})$ for the ML estimates $\hat{\boldsymbol{\beta}}$ is, for large samples, given by the inverse of this information matrix, that is,

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) \simeq \{-E[l_{ij}]\}^{-1} \equiv \mathbf{I}^{-1}(\boldsymbol{\beta}) \tag{7.1.19}$$

For example, if $k = 2$, the large-sample variance–covariance matrix is

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} V(\hat{\beta}_1) & \operatorname{cov}[\hat{\beta}_1, \hat{\beta}_2] \\ \operatorname{cov}[\hat{\beta}_1, \hat{\beta}_2] & V(\hat{\beta}_2) \end{bmatrix} \simeq - \begin{bmatrix} E[l_{11}] & E[l_{12}] \\ E[l_{12}] & E[l_{22}] \end{bmatrix}^{-1}$$

In addition, the ML estimates $\hat{\boldsymbol{\beta}}$ obtained from a stationary invertible ARMA process were shown to be asymptotically distributed as *multivariate normal* with mean vector $\boldsymbol{\beta}$ and covariance matrix $\mathbf{I}^{-1}(\boldsymbol{\beta})$ (e.g., Mann and Wald, 1943; Whittle, 1953; Hannan, 1960; Walker, 1964) in the sense that $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to the multivariate normal $N\{0, \mathbf{I}_*^{-1}(\boldsymbol{\beta})\}$ as $n \to \infty$, where $\mathbf{I}_*(\boldsymbol{\beta}) = \lim n^{-1}\mathbf{I}(\boldsymbol{\beta})$. The specific form of the information matrix $\mathbf{I}(\boldsymbol{\beta})$ and the limiting matrix $\mathbf{I}_*(\boldsymbol{\beta})$ for ARMA(p, q) models are described in Section 7.2.6, and details on the asymptotic normality of the estimator $\hat{\boldsymbol{\beta}}$ are examined for the special case of AR models in Appendix A7.5.

[5] Say over a 95% confidence region.

Now, using (7.1.5), we have

$$l_{ij} \simeq \frac{-S_{ij}}{2\sigma_a^2} \tag{7.1.20}$$

where

$$S_{ij} = \frac{\partial^2 S(\boldsymbol{\beta}|\mathbf{w})}{\partial\beta_i \partial\beta_j}$$

Furthermore, if for large samples, we approximate the expected values of l_{ij} or of S_{ij} by the values actually observed, then, using (7.1.19), we obtain

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) \simeq \{-E[l_{ij}]\}^{-1} \simeq 2\sigma_a^2\{E[S_{ij}]\}^{-1} \simeq 2\sigma_a^2\{S_{ij}\}^{-1} \tag{7.1.21}$$

Thus, for $k = 2$,

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) \simeq 2\sigma_a^2 \begin{bmatrix} \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_1^2} & \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_1\partial\beta_2} \\ \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_1\partial\beta_2} & \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_2^2} \end{bmatrix}^{-1}$$

If $S(\boldsymbol{\beta})$ were exactly quadratic in $\boldsymbol{\beta}$ over the relevant region of the parameter space, then all the derivatives S_{ij} would be constant over this region. In practice, the S_{ij} will vary somewhat, and we will usually assume that the derivatives are determined at or near the point $\hat{\boldsymbol{\beta}}$. Now, it is shown in the Appendices A7.3 and A7.4 that an estimate[6] of σ_a^2 is provided by

$$\hat{\sigma}_a^2 = \frac{S(\hat{\boldsymbol{\beta}})}{n} \tag{7.1.22}$$

and that for large samples, $\hat{\sigma}_a^2$ and $\hat{\boldsymbol{\beta}}$ are uncorrelated. Finally, the elements of (7.1.21) may be estimated from

$$\text{cov}[\hat{\beta}_i, \hat{\beta}_j] \simeq 2\hat{\sigma}_a^2 S^{ij} \tag{7.1.23}$$

where the $(k \times k)$ matrix $\{S^{ij}\}$ is given by $\{S^{ij}\} = \{S_{ij}\}^{-1}$ and the expression (7.1.23) is understood to define the variance $V(\hat{\beta}_i)$ when $j = i$.

Approximate Confidence Regions for the Parameters. In particular, these results allow us to obtain the approximate variances of our estimates. By taking the square root of these variances, we obtain approximate *standard errors* (SE) of the estimates. The standard error of an estimate $\hat{\beta}_i$ is denoted by $\text{SE}[\hat{\beta}_i]$. When we have to consider several parameters simultaneously, we need some means of judging the precision of the estimates *jointly*. One means of doing this is to determine a *confidence region*. If, for given σ_a^2, $l(\boldsymbol{\beta}, \sigma_a^2)$ is approximately quadratic in $\boldsymbol{\beta}$ in the neighborhood of $\hat{\boldsymbol{\beta}}$, then using (7.1.19) (see also

[6]Arguments can be advanced for using the divisor $n - k = n - p - q$ rather than n in (7.1.22), but for moderate-sample sizes, this modification does not make much difference.

TABLE 7.4 $S(\lambda)$ and Its First and Second Differences for Various Values of λ for Series B

$\lambda = 1 - \theta$	$S(\lambda)$	$\nabla(S)$	$\nabla^2(S)$
1.5	23,928	2,333	960
1.4	21,595	1,373	634
1.3	20,222	739	476
1.2	19,483	263	406
1.1	19,220	−143	390
1.0	19,363	−533	422
0.9	19,896	−955	508
0.8	20,851	−1,463	691
0.7	22,314	−2,154	1069
0.6	24,468	−3,223	
0.5	27,691		

Appendix A7.1), an approximate $1 - \varepsilon$ confidence region will be defined by

$$-\sum_i \sum_j E[l_{ij}](\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j) < \chi^2_\varepsilon(k) \tag{7.1.24}$$

where $\chi^2_\varepsilon(k)$ is the significance point exceeded by a proportion ε of the χ^2 distribution, having k degrees of freedom.

Alternatively, using the approximation (7.1.21) and substituting the estimate of (7.1.22) for σ_a^2, the approximate confidence region is given by[7]

$$\sum_i \sum_j S_{ij}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j) < 2\hat{\sigma}_a^2 \chi^2_\varepsilon(k) \tag{7.1.25}$$

However, for a quadratic $S(\boldsymbol{\beta})$ surface

$$S(\boldsymbol{\beta}) - S(\hat{\boldsymbol{\beta}}) = \frac{1}{2}\sum_i \sum_j S_{ij}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j) \tag{7.1.26}$$

Thus, using (7.1.22) and (7.1.25), we finally obtain the result that the approximate $1 - \varepsilon$ confidence region is bounded by the contour on the sum-of-squares surface, for which

$$S(\boldsymbol{\beta}) = S(\hat{\boldsymbol{\beta}})\left[1 + \frac{\chi^2_\varepsilon(k)}{n}\right] \tag{7.1.27}$$

Examples of the Calculation of Approximate Confidence Intervals and Regions.

1. *Example: Series B.* For Series B, values of $S(\lambda)$ and of its differences are shown in Table 7.4. The second difference of $S(\lambda)$ is not constant, and thus $S(\lambda)$ is not strictly quadratic. However, in the range from $\lambda = 0.85$ to $\lambda = 1.35$, $\nabla^2(S)$ does not change greatly, so that (7.1.27) can be expected to provide a reasonably close approx-

[7] A somewhat closer approximation based on the F distribution, which takes account of the approximate sampling distribution of $\hat{\sigma}_a^2$, may be employed. For moderate-sample sizes this refinement does not make much practical difference.

imation. With a minimum value $S(\hat{\lambda}) = 19,216$, the critical value $S(\lambda)$, defining an approximate 95% confidence interval, is then given by

$$S(\lambda) = 19,216\left(1 + \frac{3.84}{368}\right) = 19,416$$

Reading off the values of λ corresponding to $S(\lambda) = 19,416$ in Figure 7.1, we obtain an approximate confidence interval $0.98 < \lambda < 1.19$.

Alternatively, we can employ (7.1.25). Using the second difference at $\lambda = 1.1$, given in Table 7.4, to approximate the derivative, we obtain

$$S_{11} = \frac{\partial^2 S}{\partial \lambda^2} \simeq \frac{390}{(0.1)^2}$$

Also, using (7.1.22), $\hat{\sigma}_a^2 = 19,216/368 = 52.2$. Thus, the 95% confidence interval, defined by (7.1.25), is

$$\frac{390}{(0.1)^2}(\lambda - 1.09)^2 < 2 \times 52.2 \times 3.84$$

that is, $|\lambda - 1.09| < 0.10$. Thus, the interval is $0.99 < \lambda < 1.19$, which agrees closely with the previous calculation.

In this example, where there is only a single parameter λ, the use of (7.1.24) and (7.1.25) is equivalent to using an interval $\hat{\lambda} \pm u_{\varepsilon/2}\hat{\sigma}(\hat{\lambda})$, where $u_{\varepsilon/2}$ is the value, which excludes a proportion $\varepsilon/2$ in the upper tail of the standard normal distribution. An approximate standard error for $\hat{\lambda}$, $\hat{\sigma}(\hat{\lambda}) = \sqrt{2\hat{\sigma}_a^2 S_{11}^{-1}}$, is obtained from (7.1.23). In the present example,

$$V(\hat{\lambda}) = 2\hat{\sigma}_a^2 S_{11}^{-1} = \frac{2 \times 52.2 \times 0.1^2}{390} = 0.00268$$

and the approximate standard error is $\hat{\sigma}(\hat{\lambda}) = \sqrt{0.00268} = 0.052$. Thus, the approximate 95% confidence interval is $\hat{\lambda} \pm 1.96\hat{\sigma}(\hat{\lambda}) = 1.09 \pm 0.10$, as before.

Finally, we show later in Section 7.2.6 that it is possible to evaluate (7.1.19) analytically, for large samples from an MA(1) process, yielding

$$V(\hat{\lambda}) \simeq \frac{\lambda(2 - \lambda)}{n}$$

For the present example, substituting $\hat{\lambda} = 1.09$ for λ, we find that $V(\hat{\lambda}) \simeq 0.00269$, which agrees closely with the previous estimate and so yields the same standard error of 0.052 and the same confidence interval.

2. *Example: Series C.* In the identification of Series C, one model that was entertained was a (0, 2, 2) process. To illustrate the application of (7.1.27) for more than one parameter, Figure 7.5 shows an approximate 95% confidence region (shaded) for λ_0 and λ_1 of Series C. For this example, $S(\hat{\lambda}) = 4.20$, $n = 224$, and $\chi^2_{0.05}(2) = 5.99$,

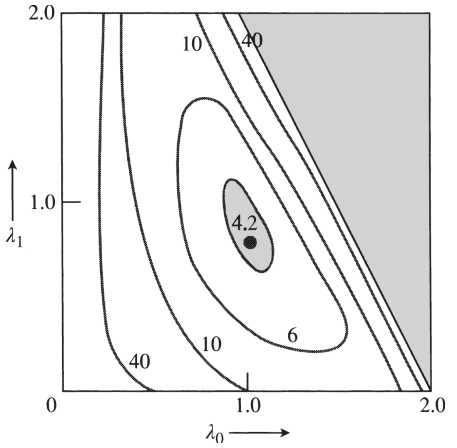

FIGURE 7.5 Sum-of-squares contours with shaded 95% confidence region for Series C, assuming a model of order (0, 2, 2).

so that the approximate 95% confidence region is bounded by the contour for which

$$S(\lambda_0, \lambda_1) = 4.20\left(1 + \frac{5.99}{224}\right) = 4.31$$

7.2 NONLINEAR ESTIMATION

7.2.1 General Method of Approach

The plotting of the sum-of-squares function is of particular importance in the study of new estimation problems because it ensures that any peculiarities in the estimation situation show up. When we are satisfied that anomalies are unlikely, other methods may be used.

We have seen that for most cases, the maximum likelihood estimates are closely approximated by the least-squares estimates, which minimize

$$S(\boldsymbol{\phi}, \theta) = \sum_{t=1}^{n} [a_t]^2 + [\mathbf{e}_*]' \boldsymbol{\Omega}^{-1} [\mathbf{e}_*]$$

and in practice, this function can be approximated by a finite sum $\sum_{t=1-Q}^{n} [a_t]^2$.

In general, considerable simplification occurs in the minimization with respect to $\boldsymbol{\beta}$, of a sum of squares $\Sigma_{t=1}^{n} [f_t(\boldsymbol{\beta})]^2$, if each $f_t(\boldsymbol{\beta})$ $(t = 1, 2, \ldots, n)$ is a *linear* function of the parameters $\boldsymbol{\beta}$. We now show that the autoregressive and moving average models differ with respect to the linearity of the $[a_t]$. For the purely autoregressive process, $[a_t] = \phi(B)[\tilde{w}_t] = [\tilde{w}_t] - \Sigma_{i=1}^{p} \phi_i [\tilde{w}_{t-i}]$ and

$$\frac{\partial [a_t]}{\partial \phi_i} = -[\tilde{w}_{t-i}] + \phi(B) \frac{\partial [\tilde{w}_t]}{\partial \phi_i}$$

Now for $u > 0, [\tilde{w}_u] = \tilde{w}_u$ and $\partial[\tilde{w}_u]/\partial\phi_i = 0$, while for $u \le 0, [\tilde{w}_u]$ and $\partial[\tilde{w}_u]/\partial\phi_i$ are both functions of $\boldsymbol{\phi}$. Thus, except for the effect of ''starting values,'' $[a_t]$ is linear in the ϕ's. By contrast, for the pure moving average process,

$$[a_t] = \theta^{-1}(B)[\tilde{w}_t] \qquad \frac{\partial[a_t]}{\partial\theta_j} = \theta^{-2}(B)[\tilde{w}_{t-j}] + \theta^{-1}(B)\frac{\partial[\tilde{w}_t]}{\partial\theta_j}$$

so that the $[a_t]$'s are always nonlinear functions of the moving average parameters.

We will see in Section 7.3 that special simplifications occur in obtaining least-squares and maximum likelihood estimates for the autoregressive process. We show in the present section how, by iterative application of linear least-squares, estimates may be obtained for any ARMA process.

Linearization of the Model. In what follows, we continue to use $\boldsymbol{\beta}$ as a general symbol for the $k = p + q$ parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$. We need, then, to minimize

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) \simeq \sum_{t=1-Q}^{n} [a_t|\tilde{\boldsymbol{w}}, \boldsymbol{\beta}]^2 = \sum_{t=1-Q}^{n} [a_t]^2$$

Expanding $[a_t]$ in a Taylor series about its value corresponding to some guessed set of parameter values $\boldsymbol{\beta}_0' = (\beta_{1,0}, \beta_{2,0}, \ldots, \beta_{k,0})$, we have approximately

$$[a_t] = [a_{t,0}] - \sum_{i=1}^{k} (\beta_i - \beta_{i,0}) x_{t,i} \tag{7.2.1}$$

where $[a_{t,0}] = [a_t|\mathbf{w}, \boldsymbol{\beta}_0]$ and

$$x_{t,i} = -\frac{\partial[a_t]}{\partial\beta_i}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$$

Now, if $\mathbf{X}$ is the $(n + Q) \times k$ matrix $\{x_{t,i}\}$, then the $n + Q$ equations (7.2.1) may be expressed as

$$[\mathbf{a}_0] = \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + [\mathbf{a}]$$

where $[\mathbf{a}_0]$ and $[\mathbf{a}]$ are column vectors with $n + Q$ elements.

The adjustments $\boldsymbol{\beta} - \boldsymbol{\beta}_0$, which minimize $S(\boldsymbol{\beta}) = S(\boldsymbol{\phi}, \boldsymbol{\theta}) = [\mathbf{a}]'[\mathbf{a}]$, may now be obtained by linear least-squares, that is, by ''regressing'' the $[a_o]$'s onto the x's. This gives the usual linear least-squares estimates, as presented in Appendix A7.2.1, of the adjustments as $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{a}_0]$, hence, $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{a}_0]$. Because the $[a_t]$'s will not be exactly linear in the parameters $\boldsymbol{\beta}$, a single adjustment will not immediately produce the final least-squares values. Instead, the adjusted values $\hat{\boldsymbol{\beta}}$ are substituted as new guesses and the process is repeated until convergence occurs. Convergence is faster if reasonably good guesses, such as may be obtained at the identification stage, are used initially. If sufficiently bad initial guesses are used, the process may not converge at all.

7.2.2 Numerical Estimates of the Derivatives

The derivatives $x_{t,i}$ may be obtained directly, as we illustrate later. They can also be computed numerically using a general nonlinear least-squares routine. This is done

by perturbing the parameters ''one at a time.'' Thus, for a given model, the values $[a_t|\mathbf{w}, \beta_{1,0}, \beta_{2,0}, \ldots, \beta_{k,0}]$ for $t = 1 - Q, \ldots, n$ are calculated recursively, using whatever preliminary ''back-forecasts'' may be needed. The calculation is then repeated for $[a_t|\mathbf{w}, \beta_{1,0} + \delta_1, \beta_{2,0}, \ldots, \beta_{k,0}]$, then for $[a_t|\mathbf{w}, \beta_{1,0}, \beta_{2,0} + \delta_2, \ldots, \beta_{k,0}]$, and so on. The negative of the required derivative is then given to sufficient accuracy using

$$x_{t,i} = \frac{[a_t|\mathbf{w}, \beta_{1,0}, \ldots, \beta_{i,0}, \ldots, \beta_{k,0}] - [a_t|\mathbf{w}, \beta_{1,0}, \ldots, \beta_{i,0} + \delta_i, \ldots, \beta_{k,0}]}{\delta_i} \tag{7.2.2}$$

The numerical method described above has the advantage of universal applicability and requires us to program the calculation of the $[a_t]$'s only, not their derivatives. General nonlinear estimation routines, which essentially require only input instructions on how to compute the $[a_t]$'s, are generally available. In some versions, it is necessary to choose the δ's in advance. In others, the program itself carries through a preliminary iteration to find suitable δ's. Many programs include special features to avoid overshoot and to speed up convergence.

Provided that the least-squares solution is not on or near a constraining boundary, the value of $\mathbf{X} = \mathbf{X}_{\hat{\boldsymbol{\beta}}}$ from the final iteration may be used to compute approximate variances, covariances, and confidence intervals. Thus, similar to the usual linear least-squares results in Appendix A7.2.3,

$$(\mathbf{X}'_{\hat{\boldsymbol{\beta}}}\mathbf{X}_{\hat{\boldsymbol{\beta}}})^{-1}\sigma_a^2$$

will approximate the variance–covariance matrix of the $\hat{\beta}$'s, and σ_a^2 will be estimated by $\hat{\sigma}_a^2 = S(\hat{\boldsymbol{\beta}})/n$.

7.2.3 Direct Evaluation of the Derivatives

We now show that it is also possible to obtain derivatives directly, but additional recursive calculations are needed. To illustrate the method, it is sufficient to consider an ARMA(1, 1) process, which can be written in either of the forms as

$$e_t = w_t - \phi w_{t+1} + \theta e_{t+1}$$
$$a_t = w_t - \phi w_{t-1} + \theta a_{t-1}$$

We have seen in Section 7.1.4, how the two versions of the model may be used in alternation, one providing initial values with which to start off a recursion with the other. We assume that a first computation has already been made yielding values of $[e_t]$, of $[a_t]$, and of $[w_0], [w_{-1}], \ldots, [w_{1-Q}]$, as in Section 7.1.5, and that $[w_{-Q}], [w_{-Q-1}], \ldots$ and hence $[a_{-Q}], [a_{-Q-1}], \ldots$ are negligible. We now show that a similar dual calculation may be used in calculating derivatives.

Using the notation $a_t^{(\phi)}$ to denote the partial derivative $\partial[a_t]/\partial\phi$, we obtain

$$e_t^{(\phi)} = w_t^{(\phi)} - \phi w_{t+1}^{(\phi)} + \theta e_{t+1}^{(\phi)} - [w_{t+1}] \tag{7.2.3}$$

$$a_t^{(\phi)} = w_t^{(\phi)} - \phi w_{t-1}^{(\phi)} + \theta a_{t-1}^{(\phi)} - [w_{t-1}] \tag{7.2.4}$$

$$e_t^{(\theta)} = w_t^{(\theta)} - \phi w_{t+1}^{(\theta)} + \theta e_{t+1}^{(\theta)} + [e_{t+1}] \tag{7.2.5}$$

$$a_t^{(\theta)} = w_t^{(\theta)} - \phi w_{t-1}^{(\theta)} + \theta a_{t-1}^{(\theta)} + [a_{t-1}] \tag{7.2.6}$$

Now,

$$\left.\begin{aligned} [w_t] &= w_t \\ w_t^{(\phi)} &= w_t^{(\theta)} = 0 \end{aligned}\right\} \qquad t = 1, 2, \ldots, n \tag{7.2.7}$$

and

$$[e_{-j}] = 0 \qquad j = 0, 1, \ldots, n \tag{7.2.8}$$

Consider equations (7.2.3) and (7.2.4). By setting $e_{n+1}^{(\phi)} = 0$ in (7.2.3), we can begin a back recursion, which using (7.2.7) and (7.2.8) eventually allows us to compute $w_{-j}^{(\phi)}$ for $j = 0, 1, \ldots, Q-1$. Since $a_{-Q}^{(\phi)}, a_{-Q-1}^{(\phi)}, \ldots$ can be taken to be zero, we can now use (7.2.4) to compute recursively the required derivatives $a_t^{(\phi)}$. In a similar way, (7.2.5) and (7.2.6) can be used to calculate the derivatives $a_t^{(\theta)}$.

7.2.4 General Least-Squares Algorithm for the Conditional Model

An approximation that we have sometimes used with long series is to set starting values for the a_t's, and hence for the derivatives in the x_t's, equal to their unconditional expectations of zero and then to proceed directly with the forward recursions. The effect is to introduce a transient into both the a_t and the x_t series, the latter being slower to die out since the x_t's depend on the a_t's. In some instances, where there is an abundance of data (say, 200 or more observations), the effect of the approximation can be nullified at the expense of some loss of information, by discarding, say, the first 10 calculated values.

If we adopt the approximation, an interesting general algorithm for this conditional model results. The ARMA(p, q) model can be written as

$$a_t = \theta^{-1}(B)\phi(B)\tilde{w}_t$$

where $w_t = \nabla^d z_t$, $\tilde{w}_t = w_t - \mu$ and

$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_i B^i - \cdots \theta_q B^q$$
$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_j B^j - \cdots \phi_p B^p$$

If the first guesses for the parameters $\boldsymbol{\beta} = (\boldsymbol{\phi}, \boldsymbol{\theta})$ are $\boldsymbol{\beta}_0 = (\boldsymbol{\phi}_0, \boldsymbol{\theta}_0)$, then

$$a_{t,0} = \theta_0^{-1}(B)\phi_0(B)\tilde{w}_t$$

and

$$-\left.\frac{\partial a_t}{\partial \phi_j}\right|_{\beta_0} = u_{t,j} = u_{t-j} \qquad -\left.\frac{\partial a_t}{\partial \theta_i}\right|_{\beta_0} = v_{t,i} = v_{t-i}$$

where

$$u_t = \theta_0^{-1}(B)\tilde{w}_t = \phi_0^{-1}(B)a_{t,0} \tag{7.2.9}$$
$$v_t = -\theta_0^{-2}(B)\phi_0(B)\tilde{w}_t = -\theta_0^{-1}(B)a_{t,0} \tag{7.2.10}$$

The a_t's, u_t's, and v_t's may be calculated recursively, with starting values for a_t's, u_t's, and v_t's set equal to zero, as follows:

$$\begin{aligned} a_{t,0} &= \tilde{w}_t - \phi_{1,0}\tilde{w}_{t-1} - \cdots - \phi_{p,0}\tilde{w}_{t-p} + \theta_{1,0}a_{t-1,0} \\ &\quad + \cdots + \theta_{q,0}a_{t-q,0} \qquad (7.2.11) \\ u_t &= \theta_{1,0}u_{t-1} + \cdots + \theta_{q,0}u_{t-q} + \tilde{w}_t \qquad (7.2.12) \\ &= \phi_{1,0}u_{t-1} + \cdots + \phi_{p,0}u_{t-p} + a_{t,0} \qquad (7.2.13) \\ v_t &= \theta_{1,0}v_{t-1} + \cdots + \theta_{q,0}v_{t-q} - a_{t,0} \qquad (7.2.14) \end{aligned}$$

Corresponding to (7.2.1), the approximate linear regression equation becomes

$$a_{t,0} = \sum_{j=1}^{p}(\phi_j - \phi_{j,0})u_{t-j} + \sum_{i=1}^{q}(\theta_i - \theta_{i,0})v_{t-i} + a_t \qquad (7.2.15)$$

The adjustments are then the regression coefficients of $a_{t,0}$ on the u_{t-j} and the v_{t-i}. By adding the adjustments to the first guesses $(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0)$, a set of ''second guesses'' are formed and these now take the place of $(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0)$ in a second iteration, in which new values of $a_{t,0}, u_t$, and v_t are computed, until convergence eventually occurs.

Alternative Form for the Algorithm. The approximate linear expansion (7.2.15) can be written in the form

$$\begin{aligned} a_{t,0} &= \sum_{j=1}^{p}(\phi_j - \phi_{j,0})B^j\phi_0^{-1}(B)a_{t,0} - \sum_{i=1}^{q}(\theta_i - \theta_{i,0})B^i\theta_0^{-1}(B)a_{t,0} + a_t \\ &= -[\phi(B) - \phi_0(B)]\phi_0^{-1}(B)a_{t,0} + [\theta(B) - \theta_0(B)]\theta_0^{-1}(B)a_{t,0} + a_t \end{aligned}$$

that is,

$$a_{t,0} = -\phi(B)[\theta_0^{-1}(B)a_{t,0}] + \theta(B)[\theta_0^{-1}(B)a_{t,0}] + a_t \qquad (7.2.16)$$

which presents the algorithm in an interesting form.

Application to an IMA(0, 2, 2) process. To illustrate the calculation with the conditional approximation, consider the estimation of least-squares values $\hat{\theta}_1, \hat{\theta}_2$ for Series C using the model of order (0, 2, 2):

$$w_t = (1 - \theta_1 B - \theta_2 B^2)a_t$$

with $w_t = \nabla^2 z_t$,

$$\begin{aligned} a_{t,0} &= w_t + \theta_{1,0}a_{t-1,0} + \theta_{2,0}a_{t-2,0} \\ v_t &= -a_{t,0} + \theta_{1,0}v_{t-1} + \theta_{2,0}v_{t-2} \end{aligned}$$

Using the initial values $\theta_{1,0} = 0.1$ and $\theta_{2,0} = 0.1$, the first adjustments to $\theta_{1,0}$ and $\theta_{2,0}$ are found by ''regressing'' $a_{t,0}$ on v_{t-1} and v_{t-2}. The process is repeated until convergence occurs. Successive parameter estimates are shown in Table 7.5.

TABLE 7.5 Convergence of Parameter Estimates for IMA(0, 2, 2) Process

Iteration	θ_1	θ_2
0	0.1000	0.1000
1	0.1247	0.1055
2	0.1266	0.1126
3	0.1286	0.1141
4	0.1290	0.1149
5	0.1292	0.1151
6	0.1293	0.1152
7	0.1293	0.1153
8	0.1293	0.1153

7.2.5 ARIMA Models Fitted to Series A–F

In Table 7.6, we summarize the models fitted by the iterative least-squares procedure of Sections 7.2.1 and 7.2.2 to Series A–F. The models fitted were identified in Chapter 6 and summarized in Tables 6.2 and 6.5. In the case of Series A, C, and D, two possible models were identified and subsequently fitted. For Series A and D, the alternative models involve the use of a stationary autoregressive operator $(1 - \phi B)$ instead of the unit-root operator $(1 - B)$. Examination of Table 7.6 shows that in both cases the autoregressive model results in a slightly smaller residual variance although the models are very similar. Even though a slightly better fit is possible with a stationary model, the IMA(0, 1, 1) model might be

TABLE 7.6 Summary of Models Fitted to Series A–F[a]

Series	Number of Observations	Fitted Models	Residual Variance[b]
A	197	$z_t - 0.92z_{t-1} = 1.45 + a_t - 0.58a_{t-1}$ (± 0.04) (± 0.08)	0.097
		$\nabla z_t = a_t - 0.70a_{t-1}$ (± 0.05)	0.101
B	369	$\nabla z_t = a_t + 0.09a_{t-1}$ (± 0.05)	52.2
C	226	$\nabla z_t - 0.82\nabla z_{t-1} = a_t$ (± 0.04)	0.018
		$\nabla^2 z_t = a_t - 0.13a_{t-1} - 0.12a_{t-2}$ (± 0.07) (± 0.07)	0.019
D	310	$z_t - 0.87z_{t-1} = 1.17 + a_t$ (± 0.03)	0.090
		$\nabla z_t = a_t - 0.06a_{t-1}$ (± 0.06)	0.096
E	100	$z_t = 14.35 + 1.42z_{t-1} - 0.73z_{t-2} + a_t$ (± 0.07) (± 0.07)	227.8
		$z_t = 11.31 + 1.57z_{t-1} - 1.02z_{t-2} + 0.21z_{t-3} + a_t$ (± 0.10) (± 0.15) (± 0.10)	218.1
F	70	$z_t = 58.87 - 0.342z_{t-1} + 0.19z_{t-2} + a_t$ (± 0.12) (± 0.12)	112.7

[a] The values $(\pm)$ under each estimate denote the standard errors of those estimates.

[b] Obtained from $S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})/n$.

preferable in these cases on the grounds that unlike the stationary model, it does not assume that the series has a fixed mean. This is especially important in predicting future values of the series. For if the level does change, a model with $d > 0$ will continue to track it, whereas a model for which $d = 0$ will be tied to a mean level that may have become out of date. It must be noted, however, that for Series D formal unit root testing to be discussed further in Section 10.1 does not support the need for differencing and suggests a preference for the stationary AR(1) model. Also, unit root testing for Series C indicates a preference for the ARIMA(1, 1, 0) model over a model in terms of second differences. Unit root testing for Series A within the ARMA(1, 1) model, though, does not reject the need for the nonstationary operator $(1 - B)$ for the autoregressive part.

The limits under the coefficients in Table 7.6 represent the standard errors of the estimates obtained from the covariance matrix $(\mathbf{X}'_{\hat{\boldsymbol{\beta}}}\mathbf{X}_{\hat{\boldsymbol{\beta}}})^{-1}\hat{\sigma}_a^2$, as described in Section 7.2.1. Note that the estimate $\hat{\phi}_3$ in the AR(3) model, fitted to the sunspot Series E, is 2.1 times its standard error, indicating that a marginally better fit is obtained by the third-order autoregressive process, as compared with the second-order autoregressive process. This is in agreement with a conclusion reached by Moran (1954).

Parameter Estimation Using R. Parameter estimation for ARIMA models based on the methods described above is available in the R software package. The relevant tools include the arima() command in the stats package and the sarima() command in the astsa package. Details of the commands are obtained by typing help(arima) and help(sarima) in R. Using the arima() command, the order of the model is specified using the argument order=c(p,d,q), and the estimation method is specified by method=c("CSS") for conditional least-squares and method=c("ML") for the full maximum likelihood method. The sarima() fits the ARIMA(p, d, q) model to a series z by maximum likelihood using the command sarima(z,p,d,q).

For illustration, we first use the arima() routine in the stats package to estimate the parameters the ARIMA(3, 0, 0) model for the sunspot data in Series E. The relevant command and a partial model output are provided below.

```
> arima(ts(seriesE),order=c(3,0,0),method=c("CSS"))

Coefficients:
         ar1      ar2     ar3  intercept
      1.5519  -1.0069  0.2076    46.7513
s.e.  0.0980   0.1540  0.0981     5.9932

sigma^2 estimated as 219.3:  log-likelihood = -411.42,  aic = NA
```

We see that the estimates of the autoregressive parameters are very close to the values provided in Table 7.6. However, using this routine, the intercept reported in the output is the mean of the series, so that the constant term in the model needs to be calculated as $\hat{\theta}_0 = \hat{\mu}(1 - \hat{\phi}_1 - \hat{\phi}_2 - \hat{\phi}_3)$. This gives an estimate for the constant of 11.57.

The commands and a partial output from performing the analysis using sarima() are as follows:

```
> library(astsa}
> sarima(ts(seriesE),3,0,0)
```

```
Coefficients:
         ar1      ar2     ar3    xmean
      1.5531  -1.0018  0.2063  48.4443
s.e.  0.0981   0.1544  0.0989   6.0706

sigma^2 estimated as 218.2:  log-likelihood=-412.49, aic 834.99
$AIC: [1] 6.465354, $AICc: [1] 6.491737, $BIC: [1] 5.569561
```

The results are close to the earlier ones. The sarima() command has an advantage in that model diagnostics of the type discussed in Chapter 8 below are provided automatically as part of the output (see, e.g., Figures 8.2 and 8.3). This allows the user to efficiently evaluate the adequacy of a fitted model and make comparisons between alternative models. For example, by fitting both the AR(2) and the AR(3) models to the sunspot series, it is readily seen that the AR(3) model provides a better fit to the data. Moreover, the fit can be improved by using a square root or log transformation of the series, although a Q–Q plot still indicates a departure from normality of the standardized residuals.

7.2.6 Large-Sample Information Matrices and Covariance Estimates

In this section, we examine in more detail the information matrix and the covariance matrix of the parameter estimates. Denote by $\mathbf{X} = [\mathbf{U} : \mathbf{V}]$, the $n \times (p+q)$ matrix of the time lagged $u_t's$ and $v_t's$ defined in (7.2.13) and (7.2.14), when the elements of $\boldsymbol{\beta}_0$ are the *true* values of the parameters, for a sample size n sufficiently large for end effects to be ignored. Then, since $x_{t,i} = -\partial[a_t]/\partial\beta_i$ and using (7.1.20),

$$E[l_{ij}] \simeq -\frac{1}{2\sigma_a^2} E\left[\frac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j}\right] = -\frac{1}{\sigma_a^2} E\left[\sum_{t=1}^{n} \frac{\partial[a_t]}{\partial\beta_i}\frac{\partial[a_t]}{\partial\beta_j}\right] = -\frac{1}{\sigma_a^2} E\left[\sum_{t=1}^{n} x_{t,i}x_{t,j}\right]$$

the information matrix for $(\boldsymbol{\phi}, \boldsymbol{\theta})$ for the mixed ARMA model is

$$\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta}) = E[\mathbf{X}'\mathbf{X}]\sigma_a^{-2} = E\begin{bmatrix}\mathbf{U}'\mathbf{U} & \mathbf{U}'\mathbf{V} \\ \mathbf{V}'\mathbf{U} & \mathbf{V}'\mathbf{V}\end{bmatrix}\sigma_a^{-2} \tag{7.2.17}$$

that is,

$$= n\sigma_a^{-2}\left[\begin{array}{ccc|ccc} \gamma_{uu}(0) & \gamma_{uu}(1) & \cdots \gamma_{uu}(p-1) & \gamma_{uv}(0) & \gamma_{uv}(-1) & \cdots \gamma_{uv}(1-q) \\ \gamma_{uu}(1) & \gamma_{uu}(0) & \cdots \gamma_{uu}(p-2) & \gamma_{uv}(1) & \gamma_{uv}(0) & \cdots \gamma_{uv}(2-q) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_{uu}(p-1) & \gamma_{uu}(p-2) & \cdots \gamma_{uu}(0) & \gamma_{uv}(p-1) & \gamma_{uv}(p-2) & \cdots \gamma_{uv}(p-q) \\ \hline \gamma_{uv}(0) & \gamma_{uv}(1) & \cdots \gamma_{uv}(p-1) & \gamma_{vv}(0) & \gamma_{vv}(1) & \cdots \gamma_{vv}(q-1) \\ \gamma_{uv}(-1) & \gamma_{uv}(0) & \cdots \gamma_{uv}(p-2) & \gamma_{vv}(1) & \gamma_{vv}(0) & \cdots \gamma_{vv}(q-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_{uv}(1-q) & \gamma_{uv}(2-q) & \cdots \gamma_{uv}(p-q) & \gamma_{vv}(q-1) & \gamma_{vv}(q-2) & \cdots \gamma_{vv}(0) \end{array}\right] \tag{7.2.18}$$

where $\gamma_{uu}(k)$ and $\gamma_{vv}(k)$ are the autocovariances for the u_t's and the v_t's, and $\gamma_{uv}(k)$ are the cross-covariances defined by

$$\gamma_{uv}(k) = \gamma_{vu}(-k) = E[u_t v_{t+k}] = E[v_t u_{t-k}]$$

The large-sample covariance matrix for the maximum likelihood estimates may be obtained using

$$\mathbf{V}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}) \simeq \mathbf{I}^{-1}(\boldsymbol{\phi}, \boldsymbol{\theta})$$

Estimates of $\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta})$ and hence of $\mathbf{V}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$ may be obtained by evaluating the u_t's and v_t's with $\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}$ and omitting the expectation sign in (7.2.17) leading to $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}) = (\boldsymbol{X}'_{\hat{\beta}} \boldsymbol{X}_{\hat{\beta}})^{-1} \hat{\sigma}_a^2$, or by substituting standard sample estimates of the autocovariances and cross-covariances in (7.2.18). Theoretical large-sample results can be obtained by noticing that, with the elements of $\boldsymbol{\beta}_0$ equal to the true values of the parameters, equations (7.2.13) and (7.2.14) imply that the derived series u_t and v_t follow *autoregressive* processes defined by

$$\phi(B) u_t = a_t \qquad \theta(B) v_t = -a_t$$

It follows that the autocovariances that appear in (7.2.18) are those for pure autoregressive processes, and the cross-covariances are the negative of those between two such processes generated by the same a_t's.

We illustrate the use of this result with a few examples.

Covariance Matrix of Parameter Estimates for AR(p) and MA(q) Processes. Let $\boldsymbol{\Gamma}_p(\boldsymbol{\phi})$ be the $p \times p$ autocovariance matrix of p successive observations from an AR(p) process with parameters $\boldsymbol{\phi}' = (\phi_1, \phi_2, \ldots, \phi_p)$. Then, using (7.2.18), the $p \times p$ covariance matrix of the estimates $\hat{\boldsymbol{\phi}}$ is given by

$$\mathbf{V}(\hat{\boldsymbol{\phi}}) \simeq n^{-1} \sigma_a^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi}) \tag{7.2.19}$$

Let $\boldsymbol{\Gamma}_q(\boldsymbol{\theta})$ be the $q \times q$ autocovariance matrix of q successive observations from an AR(q) process with parameters $\boldsymbol{\theta}' = (\theta_1, \theta_2, \ldots, \theta_q)$. Then, using (7.2.18), the $q \times q$ covariance matrix of the estimates $\hat{\boldsymbol{\theta}}$ in an MA(q) model is

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) \simeq n^{-1} \sigma_a^2 \boldsymbol{\Gamma}_q^{-1}(\boldsymbol{\theta}) \tag{7.2.20}$$

Covariances for the Zeros of an ARMA Process. It is occasionally useful to parameterize an ARMA process in terms of the zeros of $\phi(B)$ and $\theta(B)$. In this case, a particularly simple form is obtained for the covariance matrix of the parameter estimates.

Consider the ARMA(p, q) process parameterized in terms of its zeros (assumed to be real and distinct), so that

$$\prod_{i=1}^{p} (1 - G_i B) \tilde{w}_t = \prod_{j=1}^{q} (1 - H_j B) a_t$$

or

$$a_t = \prod_{i=1}^{p}(1 - G_i B) \prod_{j=1}^{q}(1 - H_j B)^{-1} \tilde{w}_t$$

The derivatives of the a_t's are then such that

$$u_{t,i} = -\frac{\partial a_t}{\partial G_i} = (1 - G_i B)^{-1} a_{t-1}$$
$$v_{t,j} = -\frac{\partial a_t}{\partial H_j} = -(1 - H_j B)^{-1} a_{t-1}$$

Hence, using (7.2.18), for large samples, the information matrix for the roots is such that

$n^{-1}\mathbf{I}(\mathbf{G}, \mathbf{H})$

$$= \left[\begin{array}{cccc|cccc} (1-G_1^2)^{-1} & (1-G_1G_2)^{-1} & \cdots & (1-G_1G_p)^{-1} & -(1-G_1H_1)^{-1} & \cdots & -(1-G_1H_q)^{-1} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ (1-G_1G_p)^{-1} & (1-G_2G_p)^{-1} & \cdots & (1-G_p^2)^{-1} & -(1-G_pH_1)^{-1} & \cdots & -(1-G_pH_q)^{-1} \\ \hline -(1-G_1H_1)^{-1} & -(1-G_2H_1)^{-1} & \cdots & -(1-G_pH_1)^{-1} & (1-H_1^2)^{-1} & \cdots & (1-H_1H_q)^{-1} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ -(1-G_1H_q)^{-1} & -(1-G_2H_q)^{-1} & \cdots & -(1-G_pH_q)^{-1} & (1-H_1H_q)^{-1} & \cdots & (1-H_q^2)^{-1} \end{array}\right] \tag{7.2.21}$$

Examples: For an AR(2) process $(1 - G_1 B)(1 - G_2 B)\tilde{w}_t = a_t$, we have

$$\mathbf{V}(\hat{G}_1, \hat{G}_2) \simeq n^{-1}\begin{bmatrix} (1-G_1^2)^{-1} & (1-G_1G_2)^{-1} \\ (1-G_1G_2)^{-1} & (1-G_2^2)^{-1} \end{bmatrix}^{-1}$$
$$= \frac{1}{n}\frac{1 - G_1G_2}{(G_1 - G_2)^2}\begin{bmatrix} (1-G_1^2)(1-G_1G_2) & -(1-G_1^2)(1-G_2^2) \\ -(1-G_1^2)(1-G_2^2) & (1-G_2^2)(1-G_1G_2) \end{bmatrix} \tag{7.2.22}$$

Exactly parallel results will be obtained for a second-order moving average process.

Similarly, for the ARMA(1,1) process $(1 - \phi B)\tilde{w}_t = (1 - \theta B)a_t$, on setting $\phi = G_1$ and $\theta = H_1$ in (7.2.21), we obtain

$$\mathbf{V}(\hat{\phi}, \hat{\theta}) \simeq n^{-1}\begin{bmatrix} (1-\phi^2)^{-1} & -(1-\phi\theta)^{-1} \\ -(1-\phi\theta)^{-1} & (1-\theta^2)^{-1} \end{bmatrix}^{-1}$$
$$= \frac{1}{n}\frac{1 - \phi\theta}{(\phi - \theta)^2}\begin{bmatrix} (1-\phi^2)(1-\phi\theta) & (1-\phi^2)(1-\theta^2) \\ (1-\phi^2)(1-\theta^2) & (1-\theta^2)(1-\phi\theta) \end{bmatrix} \tag{7.2.23}$$

The results for these two processes illustrate a duality property between the information matrices for the autoregressive model and the general ARMA(p, q) model. Namely, suppose

that the information matrix for parameters **(G, H)** of the ARMA(p, q) model

$$\prod_{i=1}^{p}(1 - G_i B)\tilde{w}_t = \prod_{j=1}^{q}(1 - H_j B)a_t$$

is denoted as $\mathbf{I}\{\mathbf{G}, \mathbf{H}|(p, q)\}$, and suppose, correspondingly, that the information matrix for the parameters **(G, H)** in the pure AR$(p + q)$ model

$$\prod_{i=1}^{p}(1 - G_i B)\prod_{j=1}^{q}(1 - H_j B)\tilde{w}_t = a_t$$

is denoted as

$$\mathbf{I}\{\mathbf{G}, \mathbf{H}|(p + q, 0)\} = \left[\begin{array}{c|c} \mathbf{I}_{GG} & \mathbf{I}_{GH} \\ \hline \mathbf{I}'_{GH} & \mathbf{I}_{HH} \end{array}\right]$$

where the matrix is partitioned after the pth row and column. Then, for moderate and large samples, we can see directly from (7.2.21) that

$$\mathbf{I}\{\mathbf{G}, \mathbf{H}|(p, q)\} \simeq \mathbf{I}\{\mathbf{G}, -\mathbf{H}|(p + q, 0)\} = \left[\begin{array}{c|c} \mathbf{I}_{GG} & -\mathbf{I}_{GH} \\ \hline -\mathbf{I}'_{GH} & \mathbf{I}_{HH} \end{array}\right] \tag{7.2.24}$$

Hence, since for moderate and large samples, the inverse of the information matrix provides a close approximation to the covariance matrix $\mathbf{V}(\hat{\mathbf{G}}, \hat{\mathbf{H}})$ of the parameter estimates, we have, correspondingly,

$$\mathbf{V}\{\hat{\mathbf{G}}, \hat{\mathbf{H}}|(p, q)\} \simeq \mathbf{V}\{\hat{\mathbf{G}}, -\hat{\mathbf{H}}|(p + q, 0)\} \tag{7.2.25}$$

7.3 SOME ESTIMATION RESULTS FOR SPECIFIC MODELS

In Appendices A7.3, A7.4, and A7.5, some estimation results for special cases are derived. These, and results obtained earlier in this chapter, are summarized here for reference.

7.3.1 Autoregressive Processes

It is possible to obtain estimates of the parameters of a pure autoregressive process by solving *certain linear* equations. We show in Appendix A7.4:

1. How exact least-squares estimates may be obtained by solving a linear system of equations (see also Section 7.5.3).
2. How, by slight modification of the coefficients in these equations, a close approximation to the exact maximum likelihood equations may be obtained.
3. How conditional least-squares estimates, as defined in Section 7.1.3, may be obtained by solving a system of linear equations of the form of the standard linear regression model normal equations.

4. How estimates that are approximations to the least-squares estimates and to the maximum likelihood estimates may be obtained using the estimated autocorrelations as coefficients in the linear Yule–Walker equations.

The estimates obtained in item 1 are, of course, identical with those given by direct minimization of $S(\boldsymbol{\phi})$, as described in general terms in Section 7.2. The estimates in 4 are the well-known approximations due to Yule and Walker. They are useful as first estimates at the identification stage but can differ appreciably from estimates 1, 2, or 3, in some cases. For instance, differences can occur for an AR(2) model if the parameter estimates $\hat{\phi}_1$ and $\hat{\phi}_2$ are highly correlated, as is the case for the AR(2) model fitted to Series E in Table 7.6.

Yule–Walker Estimates. The Yule–Walker estimates (6.3.6) are

$$\hat{\boldsymbol{\phi}} = \mathbf{R}^{-1}\boldsymbol{r}$$

where

$$\mathbf{R} = \begin{bmatrix} 1 & r_1 & \cdots & r_{p-1} \\ r_1 & 1 & \cdots & r_{p-2} \\ \vdots & \vdots & & \vdots \\ r_{p-1} & r_{p-2} & \cdots & 1 \end{bmatrix} \qquad \mathbf{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{bmatrix} \tag{7.3.1}$$

In particular, the estimates for the AR(1) and the AR(2) processes are

$$\begin{aligned} &\text{AR(1)}: \qquad \hat{\phi}_1 = r_1 \\ &\text{AR(2)}: \qquad \hat{\phi}_1 = \frac{r_1(1-r_2)}{1-r_1^2} \qquad \hat{\phi}_2 = \frac{r_2 - r_1^2}{1-r_1^2} \end{aligned} \tag{7.3.2}$$

It is shown in Appendix A7.4 that an approximation to $S(\hat{\boldsymbol{\phi}})$ is provided by

$$S(\hat{\boldsymbol{\phi}}) = \sum_{t=1}^{n} \tilde{w}_t^2 (1 - \mathbf{r}'\hat{\boldsymbol{\phi}}) \tag{7.3.3}$$

so that

$$\hat{\sigma}_a^2 = \frac{S(\hat{\boldsymbol{\phi}})}{n} = c_0(1 - \mathbf{r}'\hat{\boldsymbol{\phi}}) \tag{7.3.4}$$

where c_0 is the sample variance of the w_t's. A parallel expression relates σ_a^2 and γ_0, the theoretical variance of the w_t's [see (3.2.8)], namely,

$$\sigma_a^2 = \gamma_0(1 - \boldsymbol{\rho}'\boldsymbol{\phi})$$

where the elements of $\boldsymbol{\rho}$ and of $\boldsymbol{\phi}$ are the theoretical values. Thus, from (7.2.19) and Appendix A7.5, the covariance matrix for the estimates $\hat{\boldsymbol{\phi}}$ is

$$\boldsymbol{V}(\hat{\boldsymbol{\phi}}) \simeq n^{-1}\sigma_a^2\boldsymbol{\Gamma}^{-1} = n^{-1}(1 - \boldsymbol{\rho}'\boldsymbol{\phi})\boldsymbol{P}^{-1} \tag{7.3.5}$$

where $\boldsymbol{\Gamma}$ and $\boldsymbol{P} = (1/\gamma_0)\boldsymbol{\Gamma}$ are the autocovariance and autocorrelation matrices of p successive values of the AR(p) process.

In particular, for the AR(1) and AR(2) processes, we find that

$$\text{AR(1)}: \qquad V(\hat{\phi}) \simeq n^{-1}(1-\phi^2) \tag{7.3.6}$$

$$\text{AR(2)}: \boldsymbol{V}(\hat{\phi}_1, \hat{\phi}_2) \simeq n^{-1} \begin{bmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{bmatrix} \tag{7.3.7}$$

Estimates of the variances and covariances are obtained by substituting estimates of the parameters in (7.3.5). Thus,

$$\boldsymbol{V}(\hat{\boldsymbol{\phi}}) = n^{-1}(1 - \boldsymbol{r}'\hat{\boldsymbol{\phi}})\boldsymbol{R}^{-1} \tag{7.3.8}$$

Using (7.3.7) it is readily shown that the correlation between the estimates of the AR(2) parameters is approximately equal to $-\rho_1$. This implies, in particular, that a large lag-1 correlation in the series can give rise to unstable estimates, which may explain the differences between the Yule–Walker and the least squares estimates noted above.

7.3.2 Moving Average Processes

Maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ for moving average processes may, in simple cases, be obtained graphically, as illustrated in Section 7.1.6, or more generally, by the iterative calculation described in Section 7.2.1. From (7.2.20), it follows that for moderate and large samples, the covariance matrix for the estimates of the parameters of a qth-order moving average process is of the same form as the corresponding matrix for an autoregressive process of the same order. Thus, for the MA(1) and MA(2) processes, we find, corresponding to (7.3.6) and (7.3.7)

$$\text{MA(1)}: \qquad \mathbf{V}(\hat{\theta}) \simeq n^{-1}(1-\theta^2) \tag{7.3.9}$$

$$\text{MA(2)}: \quad \mathbf{V}(\hat{\theta}_1, \hat{\theta}_2) \simeq n^{-1} \begin{bmatrix} 1-\theta_2^2 & -\theta(1+\theta_2) \\ -\theta_1(1+\theta_2) & 1-\theta_2^2 \end{bmatrix} \tag{7.3.10}$$

7.3.3 Mixed Processes

Maximum likelihood estimates $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$ for mixed processes, as for moving average processes, may be obtained graphically in simple cases, and more generally, by iterative calculation. For moderate and large samples, the covariance matrix may be obtained by evaluating and inverting the information matrix (7.2.18). In the important special case of the ARMA(1, 1) process

$$(1-\phi B)\tilde{w}_t = (1-\theta B)a_t$$

we obtain, as in (7.2.23),

$$\mathbf{V}\,(\hat{\phi}, \hat{\theta}) \simeq n^{-1} \frac{1-\phi\theta}{(\phi-\theta)^2} \begin{bmatrix} (1-\phi^2)(1-\phi\theta) & (1-\phi^2)(1-\theta^2) \\ (1-\phi^2)(1-\theta^2) & (1-\theta^2)(1-\phi\theta) \end{bmatrix} \tag{7.3.11}$$

It is noted that when $\phi = \theta$, the variances of $\hat{\phi}$ and $\hat{\theta}$ are infinite. This is to be expected, for in this case the factor $(1-\phi B) = (1-\theta B)$ cancels on both sides of the model, which becomes

$$\tilde{w}_t = a_t$$

This is a particular case of *parameter redundancy*, which we discuss further in Section 7.3.5.

7.3.4 Separation of Linear and Nonlinear Components in Estimation

It is occasionally of interest to make an analysis in which the estimation of the parameters of the mixed model is separated into its basic linear and nonlinear parts. Consider the general mixed model $\phi(B)\tilde{w}_t = \theta(B)a_t$, which we write as $a_t = \phi(B)\theta^{-1}(B)\tilde{w}_t$, or

$$a_t = \phi(B)(\varepsilon_t|\boldsymbol{\theta}) \tag{7.3.12}$$

where

$$(\varepsilon_t|\boldsymbol{\theta}) = \theta^{-1}(B)\tilde{w}_t$$

that is,

$$\tilde{w}_t = \theta(B)(\varepsilon_t|\boldsymbol{\theta}) \tag{7.3.13}$$

For any given set of θ's, the ε_t's may be calculated recursively from (7.3.13), which may be written as

$$\varepsilon_t = \tilde{w}_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}$$

The recursion may be started by setting unknown ε_t's equal to zero. Having calculated the ε_t's, the conditional estimates $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$ may readily be obtained. These are the estimated autoregressive parameters in the linear model (7.3.12), which may be written as

$$a_t = \varepsilon_t - \phi_1\varepsilon_{t-1} - \phi_2\varepsilon_{t-2} - \cdots - \phi_p\varepsilon_{t-p} \tag{7.3.14}$$

As discussed in Section 7.3.1, the least-squares estimates of the autoregressive parameters may be found by direct solution of a set of linear equations. In simple cases, we can examine the behavior of $S(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \boldsymbol{\theta})$ and find its minimum by computing $S(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \boldsymbol{\theta})$ on a grid of $\boldsymbol{\theta}$ values and plotting contours.

Example Using Series C. One possible model for Series C considered earlier is the ARIMA(1, 1, 0) model $(1 - \phi B)w_t = a_t$ with $w_t = \nabla z_t$ and $E[w_t] = 0$. Consider now the somewhat more elaborate model $(1 - \phi B)w_t = (1 - \theta_1 B - \theta_2 B^2)a_t$. Following the argument given above, the process may be thought of as resulting from a combination of the nonlinear model $\varepsilon_t = w_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}$ and the linear model $a_t = \varepsilon_t - \phi\varepsilon_{t-1}$.

For each choice of the nonlinear parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$ within the invertibility region, a set of ε_t's was calculated recursively. Using the Yule–Walker approximation, an estimate $\hat{\phi}_{\boldsymbol{\theta}} = r_1(\varepsilon)$ could now be obtained together with

$$S(\hat{\phi}_{\boldsymbol{\theta}}, \boldsymbol{\theta}) \simeq \sum_{t=1}^{n} \varepsilon_t^2[1 - r_1^2(\varepsilon)]$$

This sum of squares was plotted for a grid of values of θ_1 and θ_2 and its contours are shown in Figure 7.6. We see that a minimum close to $\theta_1 = \theta_2 = 0$ is indicated, at which point $r_1(\varepsilon) = 0.805$. Thus, within the whole class of models of order (1, 1, 2), the simple (1, 1, 0) model $(1 - 0.8B)\nabla z_t = a_t$ is confirmed to provide an adequate representation.

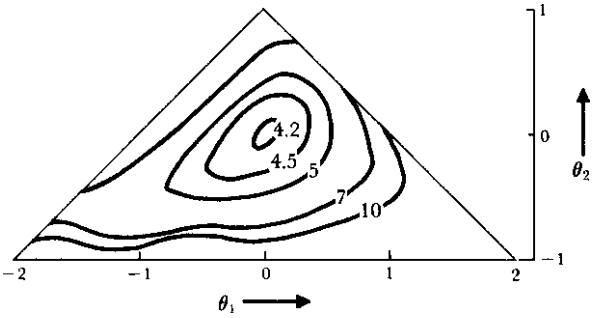

FIGURE 7.6 Counters of $S(\hat{\phi}_\theta, \theta)$ for Series C plotted over the admissible parameter space for the θ's.

7.3.5 Parameter Redundancy

The model $\phi(B)\tilde{w}_t = \theta(B)a_t$ is identical to the model

$$(1 - \alpha B)\phi(B)\tilde{w}_t = (1 - \alpha B)\theta(B)a_t$$

in which both autoregressive and moving average operators are multiplied by the same factor, $1 - \alpha B$. Serious difficulties in the estimation procedure will arise if a model is fitted that contains a redundant factor. Therefore, care is needed in avoiding the situation where redundant or near-redundant *common* factors occur. The existence of redundancy is not always obvious. For example, one can see the common factor in the ARMA(2, 1) model

$$(1 - 1.3B + 0.4B^2)\tilde{w}_t = (1 - 0.5B)a_t$$

only after factoring the left-hand side to obtain

$$(1 - 0.5B)(1 - 0.8B)\tilde{w}_t = (1 - 0.5B)a_t$$

that is, $(1 - 0.8B)\tilde{w}_t = a_t$.

In practice, it is not just exact cancellation that causes difficulties, but also near-cancellation. For example, suppose that the true model was

$$(1 - 0.4B)(1 - 0.8B)\tilde{w}_t = (1 - 0.5B)a_t \tag{7.3.15}$$

If an attempt was made to fit this model as ARMA(2, 1), extreme instability in the parameter estimates could arise because of near-cancellation of the factors $(1 - 0.4B)$ and $(1 - 0.5B)$, on the left- and right-hand sides. In this case, combinations of parameter values yielding similar $[a_t]$'s and so similar likelihoods can be found, and a change of parameter value on the left can be nearly compensated by a suitable change on the right. The sum-of-squares contour surfaces in the three-dimensional parameter space will thus approach obliquely oriented cylinders, and a line of ''near least-squares'' solutions rather than a clearly defined point minimum will be found.

From a slightly different viewpoint, we can write the model (7.3.15) in terms of an infinite autoregressive operator. Making the necessary expansion, we find that

$$(1 - 0.700B - 0.030B^2 - 0.015B^3 - 0.008B^4 - \cdots)\tilde{w}_t = a_t$$

Thus, very nearly, the model is

$$(1 - 0.7B)\tilde{w}_t = a_t \tag{7.3.16}$$

The instability of the estimates, obtained by attempting to fit an ARMA(2, 1) model, would occur because we would be trying to fit three parameters in a situation that could almost be represented by one.

A principal reason for going through the identification procedure prior to fitting the model is to avoid difficulties arising from parameter redundancy and to achieve *parsimony* in parameterization.

Redundancy in the ARMA(1,1) Model. The simplest model where the possibility occurs for direct cancellation of factors is the ARMA(1, 1) process:

$$(1 - \phi B)\tilde{w}_t = (1 - \theta B)a_t$$

In particular, if $\phi = \theta$, then whatever common value they have, $\tilde{w}_t = a_t$, so that $\tilde{w}_t$ is generated by a white noise process. The data then cannot supply information about the common parameter, and using (7.3.11), $\hat{\phi}$ and $\hat{\theta}$ have infinite variances. Furthermore, whatever the values of ϕ and θ, $S(\phi, \theta)$ must be constant on the line $\phi = \theta$. This is illustrated in Figure 7.7, which shows a sum-of-squares plot for the data of Series A. However, for these data, the least-squares values $\hat{\phi} = 0.92$ and $\hat{\theta} = 0.58$ correspond to a point that is not particularly close to the line $\phi = \theta$, and no difficulties occur in the estimation of these parameters.

In practice, if the identification technique we have recommended is adopted, these difficulties will be avoided. An ARMA(1, 1) process in which ϕ is very nearly equal to θ will normally be identified as white noise, or if the difference is nonnegligible, as an AR(1) or MA(1) process with a single small coefficient.

In summary:

1. We should avoid mixed models containing near common factors, and we should be alert to the difficulties that can result.

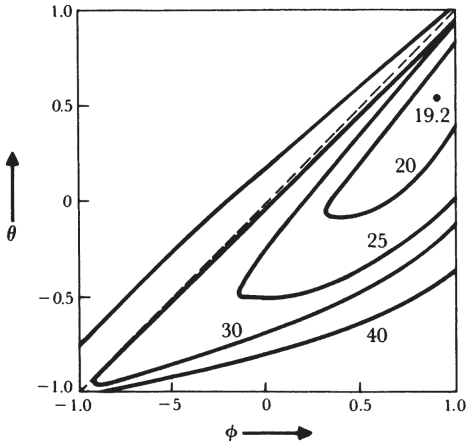

FIGURE 7.7 Sum-of-squares plot for Series A.

2. We will automatically avoid such models if we use identification and estimation procedures intelligently.

7.4 LIKELIHOOD FUNCTION BASED ON THE STATE-SPACE MODEL

In Section 5.5, we introduced the state-space model formulation of the ARMA process along with Kalman filtering and described its use for prediction. This approach also provides a convenient method to evaluate the exact likelihood function for an ARMA model. The use of this approach has been suggested by Jones (1980), Gardner et al. (1980), and others.

The state-space model form of the ARMA(p, q) model given in Section 5.5 is

$$\boldsymbol{Y}_t = \boldsymbol{\Phi}\boldsymbol{Y}_{t-1} + \boldsymbol{\Psi}a_t \quad \text{and} \quad w_t = \mathbf{H}\boldsymbol{Y}_t \tag{7.4.1}$$

where $\boldsymbol{Y}_t' = (w_t, \hat{w}_t(1), \ldots, \hat{w}_t(r-1))$, $r = \max(p, q+1)$, $\mathbf{H} = (1, 0, \ldots, 0)$,

$$\boldsymbol{\Phi} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ \phi_r & \phi_{r-1} & \cdots & \cdots & \phi_1 \end{bmatrix}$$

and $\boldsymbol{\Psi}' = (1, \psi_1, \ldots, \psi_{r-1})$. The Kalman filter equations (5.5.6)–(5.5.9) provide one-step-ahead forecasts $\hat{\boldsymbol{Y}}_{t|t-1} = E[\boldsymbol{Y}_t | w_{t-1}, \ldots, w_1]$ of the state vector $\boldsymbol{Y}_t$ and the error covariance matrix $\mathbf{V}_{t|t-1} = E[(\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1})(\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1})']$. Specifically, for the state–space form of the ARMA(p, q) model, these recursive equations are

$$\hat{\boldsymbol{Y}}_{t|t} = \hat{\boldsymbol{Y}}_{t|t-1} + \mathbf{K}_t(w_t - \hat{w}_{t|t-1}) \quad \text{with} \quad \mathbf{K}_t = \mathbf{V}_{t|t-1}\mathbf{H}'[\mathbf{H}\mathbf{V}_{t|t-1}\mathbf{H}']^{-1} \tag{7.4.2}$$

where $\hat{w}_{t|t-1} = \mathbf{H}\hat{\boldsymbol{Y}}_{t|t-1}$, and

$$\hat{\boldsymbol{Y}}_{t|t-1} = \boldsymbol{\Phi}\hat{\boldsymbol{Y}}_{t-1|t-1} \qquad \mathbf{V}_{t|t-1} = \boldsymbol{\Phi}\mathbf{V}_{t-1|t-1}\boldsymbol{\Phi}' + \sigma_a^2\boldsymbol{\Psi}\boldsymbol{\Psi}' \tag{7.4.3}$$

with

$$\mathbf{V}_{t|t} = [\mathbf{I} - \mathbf{K}_t\mathbf{H}]\mathbf{V}_{t|t-1} \tag{7.4.4}$$

for $t = 1, 2, \ldots, n$. In particular, then, the first component of the forecast vector is $\hat{w}_{t|t-1} = \mathbf{H}\hat{\boldsymbol{Y}}_{t|t-1} = E[w_t | w_{t-1}, \ldots, w_1]$, $a_{t|t-1} = w_t - \hat{w}_{t|t-1}$ is the one-step innovation, and the element $\sigma_a^2 v_t = \mathbf{H}\mathbf{V}_{t|t-1}\mathbf{H}' = E[(w_t - \hat{w}_{t|t-1})^2]$ is the one-step forecast error variance.

To obtain the exact likelihood function of the vector of n observations $\mathbf{w}' = (w_1, w_2, \ldots, w_n)$ using the above results, we note that the joint distribution of $\mathbf{w}$ can be factored as

$$p(\mathbf{w}|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = \prod_{t=1}^{n} p(w_t | w_{t-1}, \ldots, w_1; \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) \tag{7.4.5}$$

where $p(w_t|w_{t-1}, \ldots, w_1; \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2)$ denotes the conditional distribution of w_t given $w_{t-1}, \ldots, w_t$. Under normality of a_t, this conditional distribution is normal with conditional mean $\hat{w}_{t|t-1} = E[w_t|w_{t-1}, \ldots \ldots, w_1]$ and conditional variance $\sigma_a^2 v_t = E[(w_t - \hat{w}_{t|t-1})^2]$. Hence, the joint distribution of $\mathbf{w}$ can be conveniently expressed as

$$p(\mathbf{w}|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = \prod_{t=1}^{n} (2\pi\sigma_a^2 v_t)^{-1/2} \exp\left[-\frac{1}{2\sigma_a^2} \sum_{t=1}^{n} \frac{(w_t - \hat{w}_{t|t-1})^2}{v_t}\right] \tag{7.4.6}$$

where the quantities $\hat{w}_{t|t-1}$ and $\sigma_a^2 v_t$ are easily determined recursively from the Kalman filter procedure. The initial values needed to start the Kalman filter recursions are given by $\hat{\boldsymbol{Y}}_{0|0} = \mathbf{0}$, an r-dimensional vector of zeros, and $\mathbf{V}_{0|0} = \text{cov}[\boldsymbol{Y}_0]$. The elements of $\mathbf{V}_{0|0}$ can readily be determined as a function of the autocovariances γ_k and the weights ψ_k of the ARMA(p, q) process w_t, making use of the relation $w_{t+j} = \hat{w}_t(j) + \sum_{k=0}^{j-1} \psi_k a_{t+j-k}$ from Chapter 5. See Jones (1980) for further details. For example, in the case of an ARMA(1, 1) model for w_t, we have $\boldsymbol{Y}_t' = (w_t, \hat{w}_t(1))$, so

$$\mathbf{V}_{0|0} = \text{cov}[\boldsymbol{Y}_0] = \begin{bmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 - \sigma_a^2 \end{bmatrix} = \sigma_a^2 \begin{bmatrix} \sigma_a^{-2}\gamma_0 & \sigma_a^{-2}\gamma_1 \\ \sigma_a^{-2}\gamma_1 & \sigma_a^{-2}\gamma_0 - 1 \end{bmatrix}$$

It also is generally the case that the one-step-ahead forecasts $\hat{w}_{t|t-1}$ and the corresponding error variances $\sigma_a^2 v_t$ rather quickly approach their steady-state forms, in which case the Kalman filter calculations at some stage (beyond time t_0, say) could be switched to the simpler form $\hat{w}_{t|t-1} = \sum_{i=1}^{p} \phi_i w_{t-i} - \sum_{i=1}^{q} \theta_i a_{t-i|t-i-1}$, and $\sigma_a^2 v_t = \text{var}[a_{t|t-1}] = \sigma_a^2$, for $t > t_0$, where $a_{t|t-1} = w_t - \hat{w}_{t|t-1}$. For example, refer to Gardner et al. (1980) for further details. On comparison of (7.4.6) with expressions given earlier in (7.1.5) and (7.1.6), and also (A7.3.11) and (A7.3.13), the unconditional sum-of-squares function can be represented in two equivalent forms as

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} [a_t]^2 + \hat{\mathbf{e}}_*' \boldsymbol{\Omega}^{-1} \hat{\mathbf{e}}_* = \sum_{t=1}^{n} \frac{a_{t|t-1}^2}{v_t}$$

where $a_{t|t-1} = w_t - \hat{w}_{t|t-1}$, and also $|\mathbf{M}_n^{(p,q)}|^{-1} = |\boldsymbol{\Omega}||\mathbf{D}| = \prod_{t=1}^{n} v_t$.

Innovations Method. The likelihood function expressed in the form of (7.4.6) is generally referred to as the *innovations form*, and the quantities $a_{t|t-1} = w_t - \hat{w}_{t|t-1}, t = 1, \ldots, n$, are the (finite-sample) *innovations*. Calculation of the likelihood function in this form, based on the state-space representation of the ARMA process and associated Kalman filtering algorithms, has been proposed by many authors including Gardner et al. (1980), Harvey and Phillips (1979), and Jones (1980). The innovations form of the likelihood can also be obtained without directly using the state-space representation through the use of an ''innovations algorithm'' (e.g., see Ansley, 1979; Brockwell and Davis, 1991). This method essentially involves a Cholesky decomposition of an $n \times n$ band covariance matrix of the derived MA(q) process:

$$w_t' = w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

More specifically, using the notation of Appendix A7.3, we write the ARMA model relations for n observations as $\mathbf{L}_\phi \mathbf{w} = \mathbf{L}_\theta \mathbf{a} + \mathbf{F}\mathbf{e}_*$, where $\mathbf{a}' = (a_1, a_2, \ldots, a_n)$ and $\mathbf{e}_*' =$

$(w_{1-p}, \ldots, w_0, a_{1-q}, \ldots, a_0)$ is the $(p+q)$-dimensional vector of pre-sample values. Then, the covariance matrix of the vector of derived variables $\mathbf{L}_\phi \mathbf{w}$ is

$$\mathbf{\Gamma}_{w'} = \text{cov}[\mathbf{L}_\phi \mathbf{w}] = \text{cov}[\mathbf{L}_\theta \mathbf{a} + \mathbf{F}\mathbf{e}_*] = \sigma_a^2(\mathbf{L}_\theta \mathbf{L}_\theta' + \mathbf{F}\mathbf{\Omega}\mathbf{F}') \tag{7.4.7}$$

which is a *band matrix*. That is, $\mathbf{\Gamma}_{w'}$ is a matrix with nonzero elements only in a band about the main diagonal of maximum bandwidth $m = \max(p, q)$, and of bandwidth q after the first m rows since $\text{cov}[w_t', w_{t+j}'] = 0$ for $j > q$. The innovations algorithm obtains the (square-root-free) Cholesky decomposition of the band matrix $\mathbf{L}_\theta \mathbf{L}_\theta' + \mathbf{F}\mathbf{\Omega}\mathbf{F}'$ as $\mathbf{GDG}'$, where $\mathbf{G}$ is a lower triangular band matrix with bandwidth corresponding to that of $\mathbf{\Gamma}_{w'}$ and with ones on the diagonal, and $\mathbf{D}$ is a diagonal matrix with positive diagonal elements $v_t, t = 1, \ldots, n$. Hence, $\text{cov}[\mathbf{w}] = \sigma_a^2 \mathbf{L}_\phi^{-1} \mathbf{GDG}' \mathbf{L}_\phi'^{-1}$ and the quadratic form in the exponent of the likelihood function (7.4.6) is

$$\begin{aligned} \mathbf{w}'\{\text{cov}[\mathbf{w}]\}^{-1}\mathbf{w} &= \frac{1}{\sigma_a^2}\mathbf{w}'(\mathbf{L}_\phi^{-1}\mathbf{GDG}'\mathbf{L}_\phi'^{-1})^{-1}\mathbf{w} \\ &= \frac{1}{\sigma_a^2}\mathbf{e}'\mathbf{D}^{-1}\mathbf{e} = \frac{1}{\sigma_a^2}\sum_{t=1}^{n}\frac{a_{t|t-1}^2}{v_t} \end{aligned} \tag{7.4.8}$$

where $\mathbf{e} = \mathbf{G}^{-1}\mathbf{L}_\phi \mathbf{w} = (a_{1|0}, a_{2|1}, \ldots, a_{n|n-1})'$ is the vector of innovations, which are computed recursively from $\mathbf{Ge} = \mathbf{L}_\phi \mathbf{w}$. Thus, the innovations can be obtained recursively as $a_{1|0} = w_1, a_{2|1} = w_2 - \phi_1 w_1 + \theta_{1,1} a_{1|0}, \ldots, a_{m|m-1} = w_m - \sum_{i=1}^{m-1} \phi_i w_{m-i} + \sum_{i=1}^{m-1} \theta_{i,m-1} a_{m-i|m-i-1}$, and

$$a_{t|t-1} = w_t - \sum_{i=1}^{p} \phi_i w_{t-i} + \sum_{i=1}^{q} \theta_{i,t-1} a_{t-i|t-i-1} \tag{7.4.9}$$

for $t > m$, where the tth row of the matrix $\mathbf{G}$ has the form

$$[0, \ldots, 0, -\theta_{q,t-1}, \ldots, -\theta_{1,t-1}, 1, 0, \ldots, 0]$$

with the 1 in the tth (i.e., diagonal) position. In addition, the coefficients $\theta_{i,t-1}$ in (7.4.9) and the diagonal (variance) elements v_t are obtained recursively through the Cholesky decomposition procedure. In particular, the v_t are given by the recursion

$$v_t = \frac{\gamma_0(w')}{\sigma_a^2} - \sum_{j=1}^{q} \theta_{j,t-1}^2 v_{t-j} \qquad \text{for} \quad t > m \tag{7.4.10}$$

where $\gamma_0(w')/\sigma_a^2 = \text{var}[w_t']/\sigma_a^2 = 1 + \sum_{j=1}^{q} \theta_j^2$.

The ''innovations'' state-space approach to evaluating the exact likelihood function has also been shown to be quite useful in dealing with estimation problems for ARMA models when the series has missing values; see, for example, Jones (1980), Harvey and Pierse (1984), and Wincek and Reinsel (1986).

The exact likelihood function calculated using the Kalman filtering approach can be maximized using numerical optimization algorithms. These typically require the first partial derivatives of the log-likelihood with respect to the unknown parameters, and it is often beneficial to use analytical derivatives. From the form of the likelihood in (7.4.6), it is seen that this involves obtaining partial derivatives of the one-step predictions $\hat{w}_{t|t-1}$ and

of the error variances $\sigma_a^2 v_t$ for each $t = 1, \ldots, n$. Wincek and Reinsel (1986) show how the exact derivatives of $a_{t|t-1} = w_t - \hat{w}_{t|t-1}$ and $\sigma_a^2 v_t = \text{var}[a_{t|t-1}]$ with respect to the model parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ_a^2 can be obtained recursively through differentiation of the updating and prediction equations. This in turn leads to an explicit form of iterative calculations for the maximum likelihood estimation associated with the likelihood (7.4.6), similar to the nonlinear least-squares procedures detailed in Section 7.2.

7.5 ESTIMATION USING BAYES' THEOREM

7.5.1 Bayes' Theorem

In this section, we again use the symbol $\boldsymbol{\xi}$ to represent a general vector of parameters. Bayes' theorem tells us that if $p(\boldsymbol{\xi})$ is the probability distribution for $\boldsymbol{\xi}$ prior to the collection of the data, then $p(\boldsymbol{\xi}|\mathbf{z})$, the distribution of $\boldsymbol{\xi}$ posterior to the data $\mathbf{z}$, is obtained by combining the prior distribution $p(\boldsymbol{\xi})$ and the likelihood $L(\boldsymbol{\xi}|\mathbf{z})$ in the following way:

$$p(\boldsymbol{\xi}|\mathbf{z}) = \frac{p(\boldsymbol{\xi})L(\boldsymbol{\xi}|\mathbf{z})}{\int p(\boldsymbol{\xi})L(\boldsymbol{\xi}|\mathbf{z})d\boldsymbol{\xi}} \tag{7.5.1}$$

The denominator merely ensures that $p(\boldsymbol{\xi}|\mathbf{z})$ integrates to 1. The important part of the expression is the numerator, from which we see that the posterior distribution is proportional to the prior distribution multiplied by the likelihood. Savage (1962) showed that prior and posterior probabilities can be interpreted as subjective probabilities. In particular, often before the data are available, we have very little knowledge about $\boldsymbol{\xi}$, and we would be prepared to agree that over the relevant region, it would have appeared a priori just as likely that $\boldsymbol{\xi}$ had one value as another. In this case, $p(\boldsymbol{\xi})$ could be taken as *locally* uniform, and hence $p(\boldsymbol{\xi}|\mathbf{z})$ would be proportional to the likelihood.

It should be noted that for this argument to hold, it is not necessary for the prior density of $\boldsymbol{\xi}$ to be uniform over its entire range (which for some parameters could be infinite). By requiring that it be *locally uniform*, we mean that it be approximately uniform in the region in which the likelihood is appreciable and that it does not take an overwhelmingly large value outside that region.

Thus, if ξ were the weight of a chair, we could certainly say a priori that it weighed more than an ounce and less than a ton. It is also likely that when we obtained an observation z by weighing the chair on a weighing machine, which had an error standard deviation σ, we could honestly say that we would have been equally happy with a priori values in the range $z \pm 3\sigma$. The exception would be if the weighing machine said that an apparently heavy chair weighed, say, 10 ounces. In this case, the likelihood and the prior would be incompatible, and we should not, of course, use Bayes' theorem to combine them but would check the weighing machine and, if this turned out to be accurate, inspect the chair more closely.

There is, of course, some arbitrariness in this idea. Suppose that we assumed the prior distribution of ξ to be locally uniform. This then implies that the distribution of any linear function of ξ is also locally uniform. However, the prior distribution of some nonlinear transformation $\alpha = \alpha(\xi)$ (such as $\alpha = \log \xi$) could *not* be exactly locally uniform. This arbitrariness will usually have very little effect if we are able to obtain fairly precise

estimates of ξ. We will then be considering ξ only over a small range, and over such a range the transformation from ξ to, say, $\log \xi$ would often be very nearly linear.

Jeffreys (1961) has argued that it is best to choose the metric $\alpha(\xi)$ so that Fisher's measure of information $I_\alpha = -E[\partial^2 l/\partial\alpha^2]$ is independent of the value of α, and hence of ξ. This is equivalent to choosing $\alpha(\xi)$ so that the limiting variance of its maximum likelihood estimate is independent of ξ and is achieved by choosing the prior distribution of ξ to be proportional to $\sqrt{I_\xi}$.

Jeffreys justified this choice of prior on the basis of its invariance to the parameterization employed. Specifically, with this choice, the posterior distributions for $\alpha(\xi)$ and for ξ, where $\alpha(\xi)$ and ξ are connected by a one-to-one transformation, are such that $p(\xi|\mathbf{z}) = p(\alpha|\mathbf{z})\, d\alpha/d\xi$. The same result may be obtained (Box and Tiao, 1973) by the following argument. If for large samples, the expected likelihood function for $\alpha(\xi)$ approaches a normal curve, then the mean and variance of the curve summarize the information to be expected from the data. Suppose, now, that a transformation $\alpha(\xi)$ can be found in which the approximating normal curve has nearly constant variance whatever the true values of the parameter. Then, in this parameterization, the *only* information in prospect from the data is conveyed by the *location* of the expected likelihood function. To say that we know essentially nothing a priori relative to this prospective observational information is to say that we regard different *locations* of α as equally likely a priori. Equivalently, we say that α should be taken as locally uniform.

The generalization of Jeffreys' rule to deal with several parameters is that the joint prior distribution of parameters $\boldsymbol{\xi}$ be taken proportional to

$$|\mathbf{I}_\xi|^{1/2} = \left| -E\left[\frac{\partial^2 l}{\partial\xi_i \partial\xi_j}\right]\right|^{1/2} \tag{7.5.2}$$

It has been urged (e.g., Jenkins, 1964) that the likelihood itself is best considered and plotted in that metric α for which I_α is independent of α. If this is done, it will be noted that the likelihood function and the posterior density function with uniform prior are proportional.

7.5.2 Bayesian Estimation of Parameters

We now consider the estimation of the parameters in an ARIMA model from a Bayesian point of view. It is shown in Appendix A7.3 that the exact likelihood of a time series $\mathbf{z}$ of length $N = n + d$ from an ARIMA(p, d, q) process is of the form

$$L(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{z}) = \sigma_a^{-n} f(\boldsymbol{\phi}, \boldsymbol{\theta}) \exp\left[-\frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_a^2}\right] \tag{7.5.3}$$

where

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} [a_t|\mathbf{w}, \boldsymbol{\phi}, \boldsymbol{\theta}]^2 + [\mathbf{e}_*]'\boldsymbol{\Omega}^{-1}[\mathbf{e}_*] \tag{7.5.4}$$

If we have no prior information about σ_a, $\boldsymbol{\phi}$, or $\boldsymbol{\theta}$, and since information about σ_a would supply no information about $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, it is sensible, following Jeffreys, to employ a prior distribution for $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, and σ_a of the form

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a) \propto |\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta})|^{1/2} \sigma_a^{-1}$$

It follows that the posterior distribution is

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a|\mathbf{z}) \propto \sigma_a^{-(n+1)}|\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta})|^{1/2} f(\boldsymbol{\phi}, \boldsymbol{\theta}) \exp\left[-\frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_a^2}\right] \tag{7.5.5}$$

If we now integrate (7.5.5) from zero to infinity with respect to σ_a, we obtain the exact joint posterior distribution of the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ as

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{z}) \propto |\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta})|^{1/2} f(\boldsymbol{\phi}, \boldsymbol{\theta})\{S(\boldsymbol{\phi}, \boldsymbol{\theta})\}^{-n/2} \tag{7.5.6}$$

7.5.3 Autoregressive Processes

If z_t follows an ARIMA$(p, d, 0)$ process, then $w_t = \nabla^d z_t$ follows a pure AR(p) process. It is shown in Appendix A7.4 that for such a process, the factors $|\mathbf{I}(\boldsymbol{\phi})|^{1/2}$ and $f(\boldsymbol{\phi})$, which in any case are dominated by the term in $S(\boldsymbol{\phi})$, essentially cancel. This yields the remarkably simple result that given the assumptions, the parameters $\boldsymbol{\phi}$ of the AR(p) process in w_t have the posterior distribution

$$p(\boldsymbol{\phi}|\mathbf{z}) \propto \{S(\boldsymbol{\phi})\}^{-n/2} \tag{7.5.7}$$

By this argument, then, the sum-of-squares contours, which are approximate likelihood contours, are, when nothing is known a priori, also contours of posterior probability.

Joint Distribution of the Autoregressive Parameters. It is shown in Appendix A7.4 that for the pure AR process, the least-squares estimates of the ϕ's that minimize $S(\boldsymbol{\phi}) = \boldsymbol{\phi}_u' \mathbf{D} \boldsymbol{\phi}_u$ are given by

$$\hat{\boldsymbol{\phi}} = \mathbf{D}_p^{-1}\mathbf{d} \tag{7.5.8}$$

where $\boldsymbol{\phi}_u' = (1, \boldsymbol{\phi}')$,

$$\mathbf{d} = \begin{bmatrix} D_{12} \\ D_{13} \\ \vdots \\ D_{1,p+1} \end{bmatrix} \qquad \mathbf{D}_p = \begin{bmatrix} D_{22} & D_{23} & \cdots & D_{2,p+1} \\ D_{23} & D_{33} & \cdots & D_{3,p+1} \\ \vdots & \vdots & \vdots & \vdots \\ D_{2,p+1} & D_{3,p+1} & \cdots & D_{p+1,p+1} \end{bmatrix}$$

$$\mathbf{D} = \left[\begin{array}{c|c} D_{11} & -\mathbf{d}' \\ \hline -\mathbf{d} & \mathbf{D}_p \end{array}\right] \tag{7.5.9}$$

and

$$D_{ij} = D_{ji} = \tilde{w}_i\tilde{w}_j + \tilde{w}_{i+1}\tilde{w}_{j+1} + \cdots + \tilde{w}_{n+1-j}\tilde{w}_{n+1-i} \tag{7.5.10}$$

It follows that

$$S(\boldsymbol{\phi}) = \nu s_a^2 + (\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})' \mathbf{D}_p (\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}) \tag{7.5.11}$$

where

$$s_a^2 = \frac{S(\hat{\boldsymbol{\phi}})}{\nu} \qquad \nu = n - p \tag{7.5.12}$$

and

$$S(\hat{\boldsymbol{\phi}}) = \hat{\boldsymbol{\phi}}_u' \mathbf{D} \hat{\boldsymbol{\phi}}_u = D_{11} - \hat{\boldsymbol{\phi}}' \mathbf{D}_p \hat{\boldsymbol{\phi}} = D_{11} - \mathbf{d}' \mathbf{D}_p^{-1} \mathbf{d} \tag{7.5.13}$$

Thus, we can write

$$p(\boldsymbol{\phi}|\mathbf{z}) \propto \left[1 + \frac{(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})' \mathbf{D}_p (\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})}{\upsilon s_a^2} \right]^{-n/2} \tag{7.5.14}$$

Equivalently,

$$p(\boldsymbol{\phi}|\mathbf{z}) \propto \left[1 + \frac{\frac{1}{2} \sum_i \sum_j S_{ij} (\phi_i - \hat{\phi}_i)(\phi_j - \hat{\phi}_j)}{\upsilon s_a^2} \right]^{-n/2} \tag{7.5.15}$$

where

$$S_{ij} = \frac{\partial^2 S(\boldsymbol{\phi})}{\partial \phi_i \partial \phi_j} = 2 D_{i+1,j+1}$$

It follows that, a posteriori, the parameters of an autoregressive process have a multiple t distribution (A7.1.13), with $\upsilon = n - p$ degrees of freedom.

In particular, for the special case $p = 1$, $(\phi - \hat{\phi})/s_{\hat{\phi}}$ is distributed *exactly* in a Student t distribution with $n - 1$ degrees of freedom where, using the general results given above, $\hat{\phi}$ and $s_{\hat{\phi}}$ are given by

$$\hat{\phi} = \frac{D_{12}}{D_{22}} \qquad s_{\hat{\phi}} = \left[\frac{1}{n-1} \frac{D_{11}}{D_{22}} \left(1 - \frac{D_{12}^2}{D_{11} D_{22}} \right) \right]^{1/2} \tag{7.5.16}$$

The quantity $s_{\hat{\phi}}$, for large samples, tends to $[(1 - \phi^2)/n]^{1/2}$ and in the sampling theory framework is identical with the large-sample ''standard error'' for $\hat{\phi}$. However, when using this and similar expressions within the Bayesian framework, it is to be remembered that it is the parameters (ϕ in this case) that are random variables. Quantities such as $\hat{\phi}$ and $s_{\hat{\phi}}$, which are functions of data that have already occurred, are regarded as fixed.

Normal Approximation. For samples of size $n > 50$, in which we are usually interested, the normal approximation to the t distribution is adequate. Thus, very nearly, $\boldsymbol{\phi}$ has a joint p-variate normal distribution $N\{\hat{\boldsymbol{\phi}}, \mathbf{D}_p^{-1} s_a^2\}$ having mean vector $\hat{\boldsymbol{\phi}}$ and variance–covariance matrix $\mathbf{D}_p^{-1} s_a^2$.

Bayesian Regions of Highest Probability Density. In summarizing what the posterior distribution has to tell us about the probability of various $\boldsymbol{\phi}$ values, it is useful to indicate a region of *highest probability density*, called for short an HPD region (Box and Tiao, 1965). A Bayesian $1 - \varepsilon$ HPD region has the following properties:

1. Any parameter point inside the region has higher probability density than any point outside.
2. The total posterior probability mass within the region is $1 - \varepsilon$.

Since $\boldsymbol{\phi}$ has a multiple t distribution, it follows, using the result (A7.1.4), that

$$\Pr\{(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}')\mathbf{D}_p(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}) < p s_a^2 F_\varepsilon(p, \upsilon)\} = 1 - \varepsilon \tag{7.5.17}$$

defines the *exact* $1 - \varepsilon$ HPD region for $\boldsymbol{\phi}$. Now, for $\upsilon = n - p > 100$,

$$pF_\varepsilon(p, \upsilon) \simeq \chi_\varepsilon^2(p)$$

Also,

$$(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})'\mathbf{D}_p(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}) = \frac{1}{2}\sum_i \sum_j S_{ij}(\phi_i - \hat{\phi}_i)(\phi_j - \hat{\phi}_j)$$

Thus, approximately, the HPD region defined in (7.5.17) is such that

$$\sum_i \sum_j S_{ij}(\phi_i - \hat{\phi}_i)(\phi_j - \hat{\phi}_j) < 2 s_a^2 \chi_\varepsilon^2(p) \tag{7.5.18}$$

which if we set $\hat{\sigma}_a^2 = s_a^2$ is identical with the confidence region defined by (7.1.25).

Although these approximate regions are identical, it will be remembered that their interpretation is different. From a sampling theory viewpoint, we say that if a confidence region is computed according to (7.1.25), then for each of a set of repeated samples, a proportion $1 - \varepsilon$ of these regions will include the true parameter point. From the Bayesian viewpoint, we are concerned only with the single sample $\mathbf{z}$, which has actually been observed. Assuming the relevance of the noninformative prior distribution that we have taken, the HPD region includes that proportion $1 - \varepsilon$ of the resulting probability distribution of $\boldsymbol{\phi}$, given $\mathbf{z}$, which has the highest density. In other words, the probability that the value of $\boldsymbol{\phi}$, which gave rise to the data $\mathbf{z}$, lies in the HPD region is $1 - \varepsilon$.

Using (7.5.11), (7.5.12), and (7.5.18), for large samples the approximate $1 - \varepsilon$ Bayesian HPD region is bounded by a contour for which

$$S(\boldsymbol{\phi}) = S(\hat{\boldsymbol{\phi}})\left[1 + \frac{\chi_\varepsilon^2(p)}{n}\right] \tag{7.5.19}$$

which corresponds exactly with the confidence region defined by (7.1.27).

7.5.4 Moving Average Processes

If z_t follows an ARIMA$(0, d, q)$ process, then $w_t = \nabla^d z_t$ follows a pure MA(q) process. Because of the duality in estimation results and in the information matrices, in particular, between the autoregressive model and the moving average model, it follows that in the moving average case the factors $|\mathbf{I}(\boldsymbol{\theta})|^{1/2}$ and $f(\boldsymbol{\theta})$ in (7.5.6), which in any case are dominated by $S(\boldsymbol{\theta})$, also cancel for large samples. Thus, corresponding to (7.5.7), we find that the parameters $\boldsymbol{\theta}$ of the MA(q) process in w_t have the posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{z}) \propto [S(\boldsymbol{\theta})]^{-n/2} \tag{7.5.20}$$

Again the sum-of-squares contours are, for moderate samples, essentially *exact* contours of posterior density. However, because $[a_t]$ is not a linear function of the θ's, $S(\boldsymbol{\theta})$ will not be exactly quadratic in $\boldsymbol{\theta}$, though for large samples it will often be nearly so within the relevant ranges. In that case, we have approximately

$$S(\boldsymbol{\theta}) = \nu s_a^2 + \frac{1}{2}\sum_i \sum_j S_{ij}(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$

where $\upsilon s_a^2 = S(\hat{\boldsymbol{\theta}})$ and $\upsilon = n - q$. It follows, after substituting for $S(\boldsymbol{\theta})$ in (7.5.20) and using the exponential approximation, that the following holds:

1. For large samples, $\boldsymbol{\theta}$ is *approximately* distributed in a multivariate normal distribution $N\{\hat{\boldsymbol{\theta}}, 2\{S_{ij}\}^{-1}s_a^2\}$.
2. An approximate HPD region is defined by (7.5.18) or (7.5.19), with q replacing p, and $\boldsymbol{\theta}$ replacing $\boldsymbol{\phi}$.

Example: Posterior Distribution of $\lambda = 1 - \theta$ for an IMA(0, 1, 1) Process. To illustrate, Figure 7.8 shows the approximate posterior density distribution $p(\lambda|\mathbf{z})$ from the data of Series B. It is seen to be approximately normal with its mode at $\hat{\lambda} = 1.09$ and having a standard deviation of about 0.05. A 95% Bayesian HPD interval covers essentially the same range, $0.98 < \lambda < 1.19$, as did the 95% confidence interval. Note that the density has been normalized to have unit area under the curve.

7.5.5 Mixed Processes

If z_t follows an ARIMA(p, d, q) process, then $w_t = \nabla^d z_t$ follows an ARMA (p, q) process

$$\phi(B)\tilde{w}_t = \theta(B)a_t$$

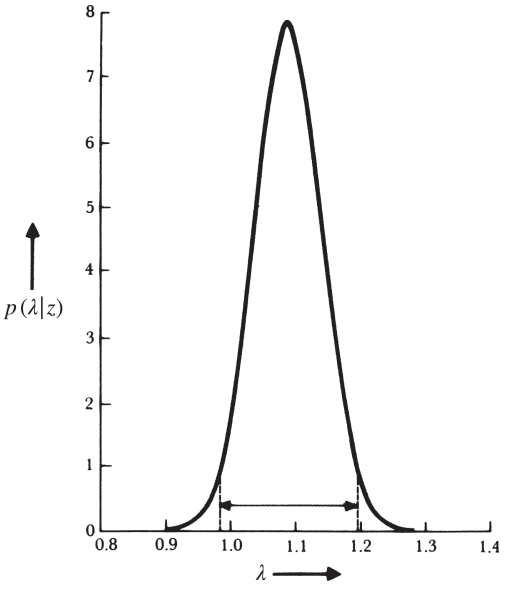

FIGURE 7.8 Posterior density $p(\lambda|\mathbf{z})$ for Series B.

It can be shown that for such a process the factors $|\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta})|^{1/2}$ and $f(\boldsymbol{\phi}, \boldsymbol{\theta})$ in (7.5.5) do not exactly cancel. Instead we can show, based on (7.2.24), that

$$|\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta})|^{1/2} f(\boldsymbol{\phi}, \boldsymbol{\theta}) = J(\boldsymbol{\phi}^* | \boldsymbol{\phi}, \boldsymbol{\theta}) \tag{7.5.21}$$

In (7.5.21), the ϕ^*'s are the $p + q$ parameters obtained by multiplying the autoregressive and moving average operators:

$$(1 - \phi_1^* B - \phi_2^* B^2 - \cdots - \phi_{p+q}^* B^{p+q}) = (1 - \phi_1 B - \cdots - \phi_p B^p) \times (1 - \theta_1 B - \cdots - \theta_q B^q)$$

and J is the Jacobian of the transformation from $\boldsymbol{\phi}^*$ to $(\boldsymbol{\phi}, \boldsymbol{\theta})$, that is,

$$p(\boldsymbol{\phi}, \boldsymbol{\theta} | \mathbf{z}) \propto J(\boldsymbol{\phi}^* | \boldsymbol{\phi}, \boldsymbol{\theta}) [S(\boldsymbol{\phi}, \boldsymbol{\theta})]^{-n/2} \tag{7.5.22}$$

In particular, for the ARMA(1, 1) process, $\phi_1^* = \phi + \theta, \phi_2^* = -\phi\theta, J = |\phi - \theta|$, and

$$p(\phi, \theta | \mathbf{z}) \propto |\phi - \theta| [S(\phi, \theta)]^{-n/2} \tag{7.5.23}$$

In this case, we see that the Jacobian will dominate in a region close to the line $\phi = \theta$ and will produce zero density on the line. This is sensible because the sum of squares $S(\phi, \theta)$ will take the finite value $\sum_{t=1}^{n} \tilde{w}_t^2$ for any $\phi = \theta$ and corresponds to our entertaining the possibility that $\tilde{w}_t$ is white noise. However, in our derivation, we have not constrained the range of the parameters. The possibility that $\phi = \theta$ is thus associated with unlimited ranges for the (equal) parameters. The effect of limiting the parameter space by, for example, introducing the requirements for stationarity and invertibility $(-1 < \phi < 1, -1 < \theta < 1)$ would be to produce a small positive value for the density, but this refinement seems scarcely worthwhile.

The Bayesian analysis reinforces the point made in Section 7.3.5 that estimation difficulties will be encountered with the mixed model and, in particular, with iterative solutions, when there is near redundancy in the parameters (i.e., near common factors between the AR and MA parts). We have already seen that the use of preliminary identification will usually ensure that these situations are avoided.

APPENDIX A7.1 REVIEW OF NORMAL DISTRIBUTION THEORY

A7.1.1 Partitioning of a Positive-Definite Quadratic Form

Consider the positive-definite quadratic form $Q_p = \mathbf{x}' \Sigma^{-1} \mathbf{x}$. Suppose that the $p \times 1$ vector $\mathbf{x}$ is partitioned after the p_1th element, so that $\mathbf{x}' = (\mathbf{x}_1' : \mathbf{x}_2') = (x_1, x_2, \ldots, x_{p_1} : x_{p_1+1}, \ldots, x_p)$, and suppose that the $p \times p$ matrix $\boldsymbol{\Sigma}$ is also partitioned after the p_1th row and column, so that

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22} \end{array}\right]$$

It is readily verified that $\boldsymbol{\Sigma}^{-1}$ can be represented as

$$\boldsymbol{\Sigma}^{-1} = \left[\begin{array}{c|c} \mathbf{I} & -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ \hline \mathbf{0} & \mathbf{I} \end{array}\right] \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \end{array}\right] \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline -\boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1} & \mathbf{I} \end{array}\right]$$

Then, since

$$\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} = (\mathbf{x}_1' \,:\, \mathbf{x}_2' - \mathbf{x}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}) \times \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \end{array}\right] \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1 \end{pmatrix}$$

$Q_p = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$ can always be written as a sum of two quadratic forms Q_{p_1} and Q_{p_2}, containing p_1 and p_2 elements, respectively, where

$$\begin{aligned} Q_p &= Q_{p_1} + Q_{p_2} \\ Q_{p_1} &= \mathbf{x}_1'\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1 \\ Q_{p_2} &= (\mathbf{x}_2 - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1)'(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}(\mathbf{x}_2 - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1) \end{aligned} \tag{A7.1.1}$$

We may also write for the determinant of $\boldsymbol{\Sigma}$

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{11}||\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}| \tag{A7.1.2}$$

A7.1.2 Two Useful Integrals

Let $\mathbf{z}'\mathbf{C}\mathbf{z}$ be a positive-definite quadratic form in $\mathbf{z}$, which has q elements, so that $\mathbf{z}' = (z_1, z_2, \ldots, z_q)$, where $-\infty < z_i < \infty, i = 1, 2, \ldots, q$, and let a, b, and m be positive real numbers. Then, it may be shown that

$$\int_R \left(a + \frac{\mathbf{z}'\mathbf{C}\mathbf{z}}{b}\right)^{-(m+q)/2} d\mathbf{z} = \frac{(b\pi)^{q/2}\Gamma(m/2)}{a^{m/2}|\mathbf{C}|^{1/2}\Gamma[(m+q)/2]} \tag{A7.1.3}$$

where the q-fold integral extends over the entire $\mathbf{z}$ space R, and

$$\frac{\int_{\mathbf{z}'\mathbf{C}\mathbf{z} > qF_0} (1 + \mathbf{z}'\mathbf{C}\mathbf{z}/m)^{-(m+q)/2} d\mathbf{z}}{\int_R (1 + \mathbf{z}'\mathbf{C}\mathbf{z}/m)^{-(m+q)/2} d\mathbf{z}} = \int_{F_0}^{\infty} p(F|q, m)\, dF \tag{A7.1.4}$$

where the function $p(F|q, m)$ is the probability density of the F distribution with q and m degrees of freedom and is defined by

$$p(F|q, m) = \frac{(q/m)^{q/2}\Gamma[(m+q)/2]}{\Gamma(q/2)\Gamma(m/2)} F^{(q-2)/2} \left(1 + \frac{q}{m}F\right)^{-(m+q)/2} \qquad F > 0 \tag{A7.1.5}$$

If m tends to infinity, then

$$\left(1 + \frac{\mathbf{z}'\mathbf{C}\mathbf{z}}{m}\right)^{-(m+q)/2} \qquad \text{tends to} \qquad e^{-(\mathbf{z}'\mathbf{C}\mathbf{z})/2}$$

and writing $qF = \chi^2$, we obtain from (A7.1.4) that

$$\frac{\int_{\mathbf{z}'\mathbf{C}\mathbf{z}>\chi_0^2} e^{-(\mathbf{z}'\mathbf{C}\mathbf{z})/2} d\mathbf{z}}{\int_R e^{-(\mathbf{z}'\mathbf{C}\mathbf{z})/2} d\mathbf{z}} = \int_{\chi_0^2}^{\infty} p(\chi^2|q) d\chi^2 \tag{A7.1.6}$$

where the function $p(\chi^2|q)$ is the probability density of the χ^2 distribution with q degrees of freedom, and is defined by

$$p(\chi^2|q) = \frac{1}{2^{q/2}\Gamma(q/2)} (\chi^2)^{(q-2)/2} e^{-\chi^2/2} \qquad \chi^2 > 0 \tag{A7.1.7}$$

Here and elsewhere $p(x)$ is used as a *general* notation to denote a probability density function for a random variable x.

A7.1.3 Normal Distribution

The random variable x is normally distributed with mean μ and standard deviation σ, or $N(\mu, \sigma^2)$, if its probability density is

$$p(x) = (2\pi)^{-1/2} (\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2} \tag{A7.1.8}$$

Thus, the unit normal variate $u = (x - \mu)/\sigma$ has a distribution $N(0, 1)$. Table E in Part Five shows ordinates $p(u_\varepsilon)$ and values u_ε such that $\Pr\{u > u_\varepsilon\} = \varepsilon$ for chosen values of ε.

Multinormal Distribution. The vector $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$ of random variables has a joint p-variate normal distribution $N\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ if its probability density function is

$$p(\mathbf{x}) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} e^{-(\mathbf{x}-\mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)/2} \tag{A7.1.9}$$

The multinormal variate $\mathbf{x}$ has mean vector $\boldsymbol{\mu} = E[\mathbf{x}]$ and variance–covariance matrix $\boldsymbol{\Sigma} = \text{cov}[\mathbf{x}]$. The probability density contours are ellipsoids defined by $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \text{constant}$. For illustration, the elliptical contours for a bivariate ($p = 2$) normal distribution are shown in Figure A7.1.

At the point $\mathbf{x} = \boldsymbol{\mu}$, the multivariate normal distribution has its maximum density

$$\max p(\mathbf{x}) = p(\boldsymbol{\mu}) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2}$$

The χ^2 Distribution as the Probability Mass Outside a Density Contour of the Multivariate Normal. For the p-variate normal distribution, (A7.1.9), the probability mass outside the density contour defined by

$$(\mathbf{x} - \mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu) = \chi_0^2$$

is given by the χ^2 integral with p degrees of freedom:

$$\int_{\chi_0^2}^{\infty} p(\chi^2|p) d\chi^2$$

where the χ^2 density function is defined as in (A7.1.7). Table F in Part Five shows values of $\chi_\varepsilon^2(p)$, such that $\Pr\{\chi^2 > \chi_\varepsilon^2(p)\} = \varepsilon$ for chosen values of ε.

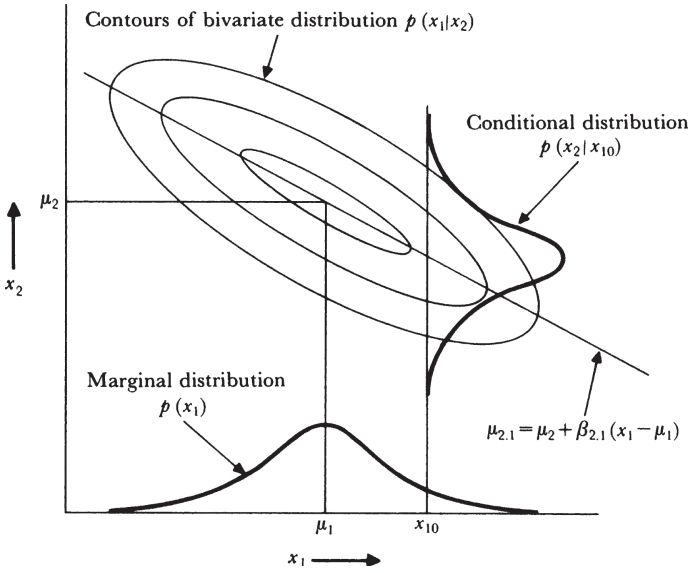

FIGURE A7.1 Contours of a bivariate normal distribution showing the marginal distribution $p(x_1)$ and the conditional distribution $p(x_2|x_{10})$ at $x_1 = x_{10}$.

Marginal and Conditional Distributions for the Multivariate Normal Distribution. Suppose that the vector of $p = p_1 + p_2$ random variables is partitioned after the first p_1 elements, so that

$$\mathbf{x}' = (\mathbf{x}_1' : \mathbf{x}_2') = (x_1, x_2, \ldots, x_{p_1} : x_{p_1+1}, \ldots, x_{p_1+p_2})$$

and that the variance–covariance matrix is

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22} \end{array}\right]$$

Then using (A7.1.1) and (A7.1.2), we can write the multivariate normal distribution for the $p = p_1 + p_2$ variates as the *marginal* distribution of $\mathbf{x}_1$ multiplied by the *conditional* distribution of $\mathbf{x}_2$ given $\mathbf{x}_1$, that is,

$$\begin{aligned} p(\mathbf{x}) &= p(\mathbf{x}_1, \mathbf{x}_2) = p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1) \\ &= (2\pi)^{-p_1/2}|\boldsymbol{\Sigma}_{11}|^{-1/2} \exp\left[-\frac{(\mathbf{x}_1 - \mu_1)'\boldsymbol{\Sigma}^{-1}{}_{11}(\mathbf{x}_1 - \mu_1)}{2}\right] \\ &\quad \times (2\pi)^{-p_2/2}|\boldsymbol{\Sigma}_{22.11}|^{-1/2} \exp\left[-\frac{(\mathbf{x}_2 - \mu_{2.1})'\boldsymbol{\Sigma}_{22.11}^{-1}(\mathbf{x}_2 - \mu_{2.1})}{2}\right] \end{aligned} \tag{A7.1.10}$$

where

$$\boldsymbol{\Sigma}_{22.11} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \tag{A7.1.11}$$

and $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \boldsymbol{\beta}_{2.1}(\mathbf{x}_1 - \boldsymbol{\mu}_1) = E[\mathbf{x}_2|\mathbf{x}_1]$ define regression hyperplanes in $(p_1 + p_2)$-dimensional space, tracing the loci of the (conditional) means of the p_2 elements of $\mathbf{x}_2$ as the p_1 elements of $\mathbf{x}_1$ vary. The $p_2 \times p_1$ matrix of regression coefficients is given by $\boldsymbol{\beta}_{2.1} = \boldsymbol{\Sigma}'_{12}\boldsymbol{\Sigma}_{11}^{-1}$.

Both marginal and conditional distributions for the multivariate normal are therefore multivariate normal distributions. It is seen that for the *multivariate normal distribution*, the conditional distribution $p(\mathbf{x}_2|\mathbf{x}_1)$ is, *except for location* (i.e., mean value), identical whatever the value of $\mathbf{x}_1$ (i.e., multivariate normal with identical variance–covariance matrix $\boldsymbol{\Sigma}_{22.11}$).

Univariate Marginals. In particular, the marginal density for a single element x_i $(i = 1, 2, \ldots, p)$ is $N(\mu_i, \sigma_i^2)$, a univariate normal with mean μ_i equal to the ith element of $\boldsymbol{\mu}$ and variance σ_i^2 equal to the ith diagonal element of $\boldsymbol{\Sigma}$.

Bivariate Normal. For illustration, the marginal and conditional distributions for a bivariate normal are shown in Figure A7.1. In this case, the marginal distribution of x_1 is $N(\mu_1, \sigma_1^2)$, while the conditional distribution of x_2 given x_1 is

$$N\left\{\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1 - \rho^2)\right\}$$

where $\rho = (\sigma_1/\sigma_2)\beta_{2.1}$ is the correlation coefficient between x_1 and x_2 and $\beta_{2.1} = \sigma_{12}/\sigma_1^2$ is the regression coefficient of x_2 on x_1.

A7.1.4 Student's *t* Distribution

The random variable x is distributed as a scaled t distribution with mean μ and scale parameter s and with ν degrees of freedom, denoted as $t(\mu, s^2, \nu)$, if its probability density is

$$p(x) = (2\pi)^{-1/2}(s^2)^{-1/2}\left(\frac{\nu}{2}\right)^{-1/2}\Gamma\left(\frac{\nu+1}{2}\right)\Gamma^{-1}\left(\frac{\nu}{2}\right)\left[1 + \frac{(x-\mu)^2}{\nu s^2}\right]^{-(\nu+1)/2} \quad \text{(A7.1.12)}$$

Thus, the standardized t variate $t = (x - \mu)/s$ has distribution $t(0, 1, \nu)$. Table G in Part Five shows values t_ε such that $\Pr\{t > t_\varepsilon\} = \varepsilon$ for chosen values of ε.

Approach to Normal Distribution. For large ν, the product

$$\left(\frac{\nu}{2}\right)^{-1/2}\Gamma\left(\frac{\nu+1}{2}\right)\Gamma^{-1}\left(\frac{\nu}{2}\right)$$

tends to unity, while the right-hand bracket in (A7.1.12) tends to $e^{-(1/2s^2)(x-\mu)^2}$. Thus, if for large ν we write $s^2 = \sigma^2$, the t distribution tends to the normal distribution (A7.1.8).

Multiple t Distribution. Let $\boldsymbol{\mu}' = (\mu_1, \mu_2, \ldots, \mu_p)$ be a $p \times 1$ vector and $\mathbf{S}$ a $p \times p$ positive-definite matrix. Then, the vector random variable $\mathbf{x}$ has a scaled multivariate t distribution $t(\boldsymbol{\mu}, \mathbf{S}, \nu)$, with mean vector $\boldsymbol{\mu}$, scaling matrix $\mathbf{S}$, and ν degrees of freedom if its probability

density is

$$p(\mathbf{x}) = (2\pi)^{-p/2}|\mathbf{S}|^{-1/2}\left(\frac{\nu}{2}\right)^{-p/2}\Gamma\left(\frac{\nu+p}{2}\right)\Gamma^{-1}\left(\frac{\nu}{2}\right) \times \left[1 + \frac{(\mathbf{x}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{\nu}\right]^{-(\nu+p)/2} \tag{A7.1.13}$$

The probability contours of the multiple t distribution are ellipsoids defined by $(\mathbf{x} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \text{constant}$.

Approach to the Multinormal Form. For large ν, the product

$$\left(\frac{\nu}{2}\right)^{-p/2}\Gamma\left(\frac{\nu+p}{2}\right)\Gamma^{-1}\left(\frac{\nu}{2}\right)$$

tends to unity; also, the right-hand bracket in (A7.1.13) tends to $e^{-(\mathbf{x}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}$. Thus, if for large υ we write $\mathbf{S} = \boldsymbol{\Sigma}$, the multiple t tends to the multivariate normal distribution (A7.1.9).

APPENDIX A7.2 REVIEW OF LINEAR LEAST-SQUARES THEORY

A7.2.1 Normal Equations and Least Squares

The linear regression model is assumed to be

$$w_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i \tag{A7.2.1}$$

where the $w_i\,(i = 1, 2, \ldots, n)$ are observations on a response or dependent variable obtained from an experiment in which the independent variables $x_{i1}, x_{i2}, \ldots, x_{ik}$ take on known *fixed* values, the β_i are unknown parameters to be estimated from the data, and the e_i are uncorrelated random errors having zero means and the same common variance σ^2.

The relations (A7.2.1) may be expressed in matrix form as

$$\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

or

$$\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{A7.2.2}$$

where the $n \times k$ matrix $\mathbf{X}$ is assumed to be of full rank k. Gauss's theorem of least-squares may be stated (Barnard, 1963) in the following form: The estimates $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$ of the parameters $\boldsymbol{\beta}$, which are linear in the observations and unbiased for $\boldsymbol{\beta}$ and which minimize the mean square error among all such estimates of any linear function $\lambda_1\beta_1 + \lambda_2\beta_2 + \cdots + \lambda_k\beta_k$ of the parameters, are obtained by minimizing the sum of squares

$$S(\boldsymbol{\beta}) = \mathbf{e}'\mathbf{e} = (\mathbf{w} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{w} - \mathbf{X}\boldsymbol{\beta}) \tag{A7.2.3}$$

To establish the minimum of $S(\boldsymbol{\beta})$, we note that the vector $\mathbf{w} - \mathbf{X}\boldsymbol{\beta}$ may be decomposed into two vectors $\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ according to

$$\mathbf{w} - \mathbf{X}\boldsymbol{\beta} = \mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{A7.2.4}$$

Hence, provided that we choose $\hat{\boldsymbol{\beta}}$ so that $\mathbf{X}'(\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$, that is,

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{w} \tag{A7.2.5}$$

it follows that

$$S(\boldsymbol{\beta}) = S(\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{A7.2.6}$$

and the vectors $\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ are orthogonal. Since the second term on the right-hand side of (A7.2.6) is a positive-definite quadratic form, it follows that the minimum is attained when $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}$$

is the *least-squares* estimate of $\boldsymbol{\beta}$ given by the solution to the *normal equation* (A7.2.5).

A7.2.2 Estimation of Error Variance

Using (A7.2.3) and (A7.2.5), the sum of squares at the minimum is

$$S(\hat{\boldsymbol{\beta}}) = (\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{w}'\mathbf{w} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \tag{A7.2.7}$$

Furthermore, if we define

$$s^2 = \frac{S(\hat{\boldsymbol{\beta}})}{n - k} \tag{A7.2.8}$$

it may be shown that $E[s^2] = \sigma^2$, and hence s^2 provides an unbiased estimate of the error variance σ^2.

A7.2.3 Covariance Matrix of Least-Squares Estimates

The covariance matrix of the least-squares estimates $\hat{\boldsymbol{\beta}}$ is defined by

$$\begin{aligned}\mathbf{V}(\hat{\boldsymbol{\beta}}) &= \text{cov}[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}'] \\ &= \text{cov}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}, \mathbf{w}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{cov}[\mathbf{w}, \mathbf{w}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\sigma^2\end{aligned} \tag{A7.2.9}$$

since $\text{cov}[\mathbf{w}, \mathbf{w}'] = \mathbf{I}\sigma^2$.

A7.2.4 Confidence Regions

Assuming normality, the quadratic forms $S(\hat{\boldsymbol{\beta}})$ and $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ in (A7.2.6) are independently distributed as σ^2 times chi-squared random variables with $n - k$ and k

degrees of freedom, respectively. Hence,

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{S(\hat{\boldsymbol{\beta}})}\frac{n-k}{k}$$

is distributed as $F(k, n-k)$. Using (A7.2.8), it follows that

$$(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \leq ks^2F_\varepsilon(k, n-k) \tag{A7.2.10}$$

defines a $1-\varepsilon$ confidence region for $\boldsymbol{\beta}$.

A7.2.5 Correlated Errors

Suppose that the errors $\mathbf{e}$ in (A7.2.2) have a *known* covariance matrix $\mathbf{V}$, and let $\mathbf{P}$ be an $n \times n$ nonsingular matrix such that $\mathbf{V}^{-1} = \mathbf{PP}'/\sigma^2$, so that $\mathbf{P}'\mathbf{VP} = \mathbf{I}\sigma^2$. Then, (A7.2.2) may be transformed into

$$\mathbf{P}'\mathbf{w} = \mathbf{P}'\mathbf{X}\boldsymbol{\beta} + \mathbf{P}'\mathbf{e}$$

or

$$\mathbf{w}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{e}^* \tag{A7.2.11}$$

where $\mathbf{w}^* = \mathbf{P}'\mathbf{w}$ and $\mathbf{X}^* = \mathbf{P}'\mathbf{X}$. The covariance matrix of $\mathbf{e}^* = \mathbf{P}'\mathbf{e}$ in (A7.2.11) is

$$\text{cov}[\mathbf{P}'\mathbf{e}, \mathbf{e}'\mathbf{P}] = \mathbf{P}'\text{cov}[\mathbf{e}, \mathbf{e}']\mathbf{P} = \mathbf{P}'\mathbf{VP} = \mathbf{I}\sigma^2$$

Hence, we may apply ordinary least-squares theory with $\mathbf{V} = \mathbf{I}\sigma^2$ to the *transformed* model (A7.2.11), in which $\mathbf{w}$ is replaced by $\mathbf{w}^* = \mathbf{P}'\mathbf{w}$ and $\mathbf{X}$ by $\mathbf{X}^* = \mathbf{P}'\mathbf{X}$. Thus, we obtain the estimates

$$\hat{\boldsymbol{\beta}}_G = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{w}^*$$

with $\mathbf{V}(\hat{\boldsymbol{\beta}}_G) = \text{cov}[\hat{\boldsymbol{\beta}}_\mathbf{G}] = \sigma^2(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}$. In terms of the *original* variables $\mathbf{X}$ and $\mathbf{w}$ of the regression model, since $\mathbf{PP}' = \sigma^2\mathbf{V}^{-1}$, the estimate is

$$\hat{\boldsymbol{\beta}}_G = (\mathbf{X}'\mathbf{PP}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{PP}'\mathbf{w} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{w} \tag{A7.2.12}$$

with

$$\mathbf{V}(\hat{\boldsymbol{\beta}}_G) = \text{cov}[\hat{\boldsymbol{\beta}}_G] = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

The estimator $\hat{\boldsymbol{\beta}}_G$ in (A7.2.12) is generally referred to as the *generalized least-squares* (GLS) estimator, and it follows that this is the estimate of $\boldsymbol{\beta}$ obtained by minimizing the *generalized* sum of squares function

$$S(\boldsymbol{\beta}|\mathbf{V}) = (\mathbf{w}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{w}-\mathbf{X}\boldsymbol{\beta})$$

APPENDIX A7.3 EXACT LIKELIHOOD FUNCTION FOR MOVING AVERAGE AND MIXED PROCESSES

To obtain the required likelihood function for an MA(q) model, we have to derive the probability density function for a series $\mathbf{w}' = (w_1, w_2, \ldots, w_n)$ assumed to be generated by an invertible moving average model of order q:

$$\tilde{w}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \tag{A7.3.1}$$

where $\tilde{w}_t = w_t - \mu$, with $\mu = E[w_t]$. Under the assumption that the a_t's and the $\tilde{w}_t$'s are normally distributed, the joint density may be written as

$$p(\mathbf{w}|\boldsymbol{\theta}, \sigma_a^2, \mu) = (2\pi\sigma_a^2)^{-n/2} |\mathbf{M}_n^{(0,q)}|^{1/2} \exp\left[\frac{-\tilde{\mathbf{w}}'\mathbf{M}_n^{(0,q)}\tilde{\mathbf{w}}}{2\sigma_a^2}\right] \tag{A7.3.2}$$

where $(\mathbf{M}_n^{(p,q)})^{-1}\sigma_a^2$ denotes the $n \times n$ covariance matrix of the w_t's for an ARMA(p, q) process. We now consider a convenient way of evaluating $\tilde{\mathbf{w}}'\mathbf{M}_n^{(0,q)}\tilde{\mathbf{w}}$, and for simplicity, we suppose that $\mu = 0$, so that $w_t = \tilde{w}_t$.

Using the model (A7.3.1), we can write down the n equations:

$$w_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \qquad (t = 1, 2, \ldots, n)$$

These n equations can be conveniently expressed in matrix form in terms of the n-dimensional vectors $\mathbf{w}' = (w_1, w_2, \ldots, w_n)$ and $\mathbf{a}' = (a_1, a_2, \ldots, a_n)$, and the q-dimensional vector of preliminary values $\mathbf{a}_*' = (a_{1-q}, a_{2-q}, \ldots, a_0)$ as

$$\mathbf{w} = \mathbf{L}_\theta \mathbf{a} + \mathbf{F}\mathbf{a}_*$$

where $\mathbf{L}_\theta$ is an $n \times n$ lower triangular matrix with l's on the leading diagonal, $-\theta_1$ on the first subdiagonal, $-\theta_2$ on the second subdiagonal, and so on, with $\theta_i = 0$ for $i > q$. Further, $\mathbf{F}$ is an $n \times q$ matrix with the form $\mathbf{F} = (\mathbf{B}_q', \mathbf{0}')'$ where $\mathbf{B}_q$ is $q \times q$ equal to

$$\mathbf{B}_q = -\begin{bmatrix} \theta_q & \theta_{q-1} & \ldots & \theta_1 \\ 0 & \theta_q & \ldots & \theta_2 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \theta_q \end{bmatrix}$$

Now the joint distribution of the $n + q$ values, which are the elements of $(\mathbf{a}', \mathbf{a}_*')$, is

$$p(\mathbf{a}, \mathbf{a}_*|\sigma_a^2) = (2\pi\sigma_a^2)^{-(n+q)/2} \exp\left[-\frac{1}{2\sigma_a^2}(\mathbf{a}'\mathbf{a} + \mathbf{a}_*'\mathbf{a}_*)\right]$$

Noting that the transformation from $(\mathbf{a}, \mathbf{a}_*)$ to $(\mathbf{w}, \mathbf{a}_*)$ has unit Jacobian and $\mathbf{a} = \mathbf{L}_\theta^{-1}(\mathbf{w} - \mathbf{F}\mathbf{a}_*)$, the joint distribution of $\mathbf{w} = \mathbf{L}_\theta\mathbf{a} + \mathbf{F}\mathbf{a}_*$ and $\mathbf{a}_*$ is

$$p(\mathbf{w}, \mathbf{a}_*|\boldsymbol{\theta}, \sigma_a^2) = (2\pi\sigma_a^2)^{-(n+q)/2} \exp\left[-\frac{1}{2\sigma_a^2} S(\boldsymbol{\theta}, \mathbf{a}_*)\right]$$

where

$$S(\boldsymbol{\theta}, \mathbf{a}_*) = (\mathbf{w} - \mathbf{F}\mathbf{a}_*)' \mathbf{L}_\theta'^{-1} \mathbf{L}_\theta^{-1} (\mathbf{w} - \mathbf{F}\mathbf{a}_*) + \mathbf{a}_*' \mathbf{a}_* \tag{A7.3.3}$$

Now, let $\hat{\mathbf{a}}_*$ be the vector of values that minimize $S(\boldsymbol{\theta}, \mathbf{a}_*)$, which from generalized least-squares theory can be shown to equal $\hat{\mathbf{a}}_* = \mathbf{D}^{-1}\mathbf{F}'\mathbf{L}_\theta'^{-1}\mathbf{L}_\theta^{-1}\mathbf{w}$, where $\mathbf{D} = \mathbf{I}_q + \mathbf{F}'\mathbf{L}_\theta'^{-1}\mathbf{L}_\theta^{-1}\mathbf{F}$. Then, using the result (A7.2.6), we have

$$S(\boldsymbol{\theta}, \mathbf{a}_*) = S(\boldsymbol{\theta}) + (\mathbf{a}_* - \hat{\mathbf{a}}_*)' \mathbf{D} (\mathbf{a}_* - \hat{\mathbf{a}}_*)$$

where

$$S(\boldsymbol{\theta}) = S(\boldsymbol{\theta}, \hat{\mathbf{a}}_*) = (\mathbf{w} - \mathbf{F}\hat{\mathbf{a}}_*)' \mathbf{L}_\theta'^{-1} \mathbf{L}_\theta^{-1} (\mathbf{w} - \mathbf{F}\hat{\mathbf{a}}_*) + \hat{\mathbf{a}}_*' \hat{\mathbf{a}}_* \tag{A7.3.4}$$

is a function of the observations $\mathbf{w}$ but not of the preliminary values $\mathbf{a}_*$. Thus,

$$p(\mathbf{w}, \mathbf{a}_* | \boldsymbol{\theta}, \sigma_a^2) = (2\pi\sigma_a^2)^{-(n+q)/2} \exp \left\{ -\frac{1}{2\sigma_a^2} [S(\boldsymbol{\theta}) + (\mathbf{a}_* - \hat{\mathbf{a}}_*)' \mathbf{D} (\mathbf{a}_* - \hat{\mathbf{a}}_*)] \right\}$$

However, since the joint distribution of $\mathbf{w}$ and $\mathbf{a}_*$ can be factored as

$$p(\mathbf{w}, \mathbf{a}_* | \boldsymbol{\theta}, \sigma_a^2) = p(\mathbf{w} | \boldsymbol{\theta}, \sigma_a^2) p(\mathbf{a}_* | \mathbf{w}, \boldsymbol{\theta}, \sigma_a^2)$$

it follows, similar to (A7.1.10), that

$$p(\mathbf{a}_* | \mathbf{w}, \boldsymbol{\theta}, \sigma_a^2) = (2\pi\sigma_a^2)^{-q/2} \mid \mathbf{D} \mid^{1/2} \exp \left[-\frac{1}{2\sigma_a^2} (\mathbf{a}_* - \hat{\mathbf{a}}_*)' \mathbf{D} (\mathbf{a}_* - \hat{\mathbf{a}}_*) \right] \tag{A7.3.5}$$

$$p(\mathbf{w} | \boldsymbol{\theta}, \sigma_a^2) = (2\pi\sigma_a^2)^{-n/2} \mid \mathbf{D} \mid^{-1/2} \exp \left[-\frac{1}{2\sigma_a^2} S(\boldsymbol{\theta}) \right] \tag{A7.3.6}$$

We can now deduce the following:

1. From (A7.3.5), we see that $\hat{\mathbf{a}}_*$ is the conditional expectation of $\mathbf{a}_*$ given $\mathbf{w}$ and $\boldsymbol{\theta}$. Thus, using the notation introduced in Section 7.1.4, we obtain

$$\hat{\mathbf{a}}_* = [\mathbf{a}_* | \mathbf{w}, \boldsymbol{\theta}] = [\mathbf{a}_*]$$

where $[\mathbf{a}] = \mathbf{L}_\theta^{-1}(\mathbf{w} - \mathbf{F}[\mathbf{a}_*])$ is the conditional expectation of $\mathbf{a}$ given $\mathbf{w}$ and $\boldsymbol{\theta}$, and using (A7.3.4):

$$S(\boldsymbol{\theta}) = [\mathbf{a}]'[\mathbf{a}] + [\mathbf{a}_*]'[\mathbf{a}_*] = \sum_{t=1-q}^{n} [a_t]^2 \tag{A7.3.7}$$

To compute $S(\boldsymbol{\theta})$, the quantities $[a_t] = [a_t | \mathbf{w}, \boldsymbol{\theta}]$ may be obtained by using the estimates $[\mathbf{a}_*]' = ([a_{1-q}], [a_{2-q}], \ldots, [a_0])$ obtained as above by back-forecasting for preliminary values, and computing the elements $[a_1], [a_2], \ldots, [a_n]$ of $[\mathbf{a}]$ recursively from the relation $\mathbf{L}_\theta[\mathbf{a}] = \mathbf{w} - \mathbf{F}[\mathbf{a}_*]$ as

$$[a_t] = w_t + \theta_1 [a_{t-1}] + \ldots + \theta_q [a_{t-q}] \qquad (t = 1, 2, \ldots, n)$$

Note that if the expression for $\hat{\mathbf{a}}_*$ is utilized in (A7.3.4), after rearranging we obtain

$$S(\boldsymbol{\theta}) = \mathbf{w}'\mathbf{L}_\theta'^{-1}(\mathbf{I}_n - \mathbf{L}_\theta^{-1}\mathbf{F}\mathbf{D}^{-1}\mathbf{F}'\mathbf{L}_\theta'^{-1})\mathbf{L}_\theta^{-1}\mathbf{w} = \mathbf{a}^{0\prime}\mathbf{a}^0 - \hat{\mathbf{a}}_*'\mathbf{D}\hat{\mathbf{a}}_*$$

where $\mathbf{a}^0 = \mathbf{L}_\theta^{-1}\mathbf{w}$ denotes the vector whose elements a_t^0 can be calculated recursively from $a_t^0 = w_t + \theta_1 a_{t-1}^0 + \cdots + \theta_q a_{t-q}^0$, $t = 1, 2, \ldots, n$, by setting the initial values $\mathbf{a}_*$ equal to zero. Hence, the first term described above, $S_*(\boldsymbol{\theta}) = \mathbf{a}^{0\prime}\mathbf{a}^0 = \sum_{t=1}^n (a_t^0)^2$, is the conditional sum-of-squares function, given $\mathbf{a}_* = \mathbf{0}$, as discussed in Section 7.1.2.

2. In addition, we find that

$$\mathbf{M}_n^{(0,q)} = \mathbf{L}_\theta'^{-1}(\mathbf{I}_n - \mathbf{L}_\theta^{-1}\mathbf{F}\mathbf{D}^{-1}\mathbf{F}'\mathbf{L}_\theta'^{-1})\mathbf{L}_\theta^{-1}$$

and $S(\boldsymbol{\theta}) = \mathbf{w}'\mathbf{M}_n^{(0,q)}\mathbf{w}$. Also, by comparing (A7.3.6) and (A7.3.2), we have

$$|\mathbf{D}|^{-1} = |\mathbf{M}_n^{(0,q)}|$$

3. The back-forecasts $\hat{\mathbf{a}}_* = [\mathbf{a}_*]$ can be calculated most conveniently from $\hat{\mathbf{a}}_* = \mathbf{D}^{-1}\mathbf{F}'\mathbf{u}$ (i.e., by solving $\mathbf{D}\hat{\mathbf{a}}_* = \mathbf{F}'\mathbf{u}$), where $\mathbf{u} = \mathbf{L}_\theta'^{-1}\mathbf{L}_\theta^{-1}\mathbf{w} = \mathbf{L}_\theta'^{-1}\mathbf{a}^0 = (u_1, u_2, \ldots, u_n)'$. Note that the elements u_t of $\mathbf{u}$ are calculated through a backward recursion as

$$u_t = a_t^0 + \theta_1 u_{t+1} + \ldots + \theta_q u_{t+q}$$

from $t = n$ down to $t = 1$, using zero starting values $u_{n+1} = \ldots = u_{n+q} = 0$, where the a_t^0 denote the estimates of the a_t conditional on the zero starting values $\mathbf{a}_* = \mathbf{0}$. Also, the vector $\mathbf{h} = \mathbf{F}'\mathbf{u}$ consists of the elements $h_j = -\sum_{i=1}^j \theta_{q-j+i}u_i$, $j = 1, \ldots, q$.

4. Finally, using (A7.3.6) and (A7.3.7), the unconditional likelihood is given exactly by

$$L(\boldsymbol{\theta}, \sigma_a^2|\mathbf{w}) = (\sigma_a^2)^{-n/2}|\mathbf{D}|^{-1/2}\exp\left\{-\frac{1}{2\sigma_a^2}\sum_{t=1-q}^{n}[a_t]^2\right\} \tag{A7.3.8}$$

For example, in the MA(1) model with $q = 1$, we have $\mathbf{F}' = -(\theta, 0, \ldots, 0)$, an n-dimensional vector, and $\mathbf{L}_\theta$ is such that $\mathbf{L}_\theta^{-1}$ has first column equal to $(1, \theta, \theta^2, \ldots, \theta^{n-1})'$, so that

$$\mathbf{D} = 1 + \mathbf{F}'\mathbf{L}_\theta'^{-1}\mathbf{L}_\theta^{-1}\mathbf{F} = 1 + \theta^2 + \theta^4 + \ldots + \theta^{2n} = \frac{1 - \theta^{2(n+1)}}{1 - \theta^2}$$

In addition, the conditional values a_t^0 are computed recursively as $a_t^0 = w_t + \theta a_{t-1}^0$, $t = 1, 2, \ldots, n$, using the zero initial value $a_0^0 = 0$, and the values of the vector $\mathbf{u} = \mathbf{L}_\theta'^{-1}\mathbf{a}^0$ are computed in the backward recursion as $u_t = a_t^0 + \theta u_{t+1}$, from $t = n$ to $t = 1$, with $u_{n+1} = 0$. Then,

$$\hat{\mathbf{a}}_* = [a_0] = -\mathbf{D}^{-1}\theta u_1 = -\mathbf{D}^{-1}u_0 = -\frac{u_0(1 - \theta^2)}{1 - \theta^{2(n+1)}}$$

where $u_0 = a_0^0 + \theta u_1 = \theta u_1$, and the exact likelihood for the MA(1) process is

$$L(\theta, \sigma_a^2|\mathbf{w}) = (\sigma_a^2)^{-n/2} \frac{(1-\theta^2)^{1/2}}{(1-\theta^{2(n+1)})^{1/2}} \exp\left\{ -\frac{1}{2\sigma_a^2} \sum_{t=0}^{n} [a_t]^2 \right\} \tag{A7.3.9}$$

Extension to the Autoregressive and Mixed Processes. The method outlined above may be readily extended to provide the unconditional likelihood for the general mixed model

$$\phi(B)\tilde{w}_t = \theta(B)a_t \tag{A7.3.10}$$

which, with $w_t = \nabla^d z_t$, defines the general ARIMA process. Details of the derivation have been presented by Newbold (1974) and Ljung and Box (1979), while an alternative approach to obtain the exact likelihood that uses the Cholesky decomposition of a band covariance matrix (i.e., the innovations method as discussed in Section 7.4) was given by Ansley (1979). First, assuming a zero mean for the process, the relations for the ARMA model may be written in matrix form, similar to before, as

$$\mathbf{L}_\phi \mathbf{w} = \mathbf{L}_\theta \mathbf{a} + \mathbf{F}\mathbf{e}_*$$

where $\mathbf{L}_\phi$ is an $n \times n$ matrix of the same form as $\mathbf{L}_\theta$ but with ϕ_i's in place of θ_i's, $\mathbf{e}_*' = (\mathbf{w}_*', \mathbf{a}_*') = (w_{1-p}, \ldots, w_0, a_{1-q}, \ldots, a_0)$ is the $(p+q)$-dimensional vector of initial values, and

$$\mathbf{F} = \begin{bmatrix} \mathbf{A}_p & \mathbf{B}_q \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

with

$$\mathbf{A}_p = \begin{bmatrix} \phi_p & \phi_{p-1} & \cdots & \phi_1 \\ 0 & \phi_p & \cdots & \phi_2 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \phi_p \end{bmatrix} \quad \text{and} \quad \mathbf{B}_q = -\begin{bmatrix} \theta_q & \theta_{q-1} & \cdots & \theta_1 \\ 0 & \theta_q & \cdots & \theta_2 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \theta_q \end{bmatrix}$$

Let $\mathbf{\Omega}\sigma_a^2 = E[\mathbf{e}_*\mathbf{e}_*']$ denote the covariance matrix of $\mathbf{e}_*$. This matrix has the form

$$\mathbf{\Omega}\sigma_a^2 = \begin{bmatrix} \sigma_a^{-2}\mathbf{\Gamma}_p & \mathbf{C}' \\ \mathbf{C} & \mathbf{I}_q \end{bmatrix} \sigma_a^2$$

where $\mathbf{\Gamma}_p = E[\mathbf{w}_*\mathbf{w}_*']$ is a $p \times p$ matrix with (i,j)th element γ_{i-j}, and $\sigma_a^2\mathbf{C} = E[\mathbf{a}_*\mathbf{w}_*']$ has elements defined by $E[a_{i-q}w_{j-p}] = \sigma_a^2\psi_{j-i-p+q}$ for $j - i - p + q \geq 0$ and 0 otherwise. The ψ_k are the coefficients in the infinite MA operator $\psi(B) = \phi^{-1}(B)\theta(B) = \sum_{k=0}^{\infty} \psi_k B^k, \psi_0 = 1$, and are easily determined recursively through equations in Section 3.4. The autocovariances γ_k in $\mathbf{\Gamma}_p$ can directly be determined in terms of the coefficients ϕ_i, θ_i, and σ_a^2, through use of the first $(p+1)$ equations (3.4.2) (see, e.g., Ljung and Box, 1979).

Similar to the result in (A7.3.3), since $\mathbf{a} = \mathbf{L}_\theta^{-1}(\mathbf{L}_\phi \mathbf{w} - \mathbf{F}\mathbf{e}_*)$ and $\mathbf{e}_*$ are independent, the joint distribution of $\mathbf{w}$ and $\mathbf{e}_*$ is

$$p(\mathbf{w}, \mathbf{e}_*|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = (2\pi\sigma_a^2)^{-(n+p+q)/2} \, |\boldsymbol{\Omega}|^{-1/2} \exp\left[-\frac{1}{2\sigma_a^2} S(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{e}_*)\right]$$

where

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{e}_*) = (\mathbf{L}_\phi \mathbf{w} - \mathbf{F}\mathbf{e}_*)' \mathbf{L}_\theta'^{-1} \mathbf{L}_\theta^{-1} (\mathbf{L}_\phi \mathbf{w} - \mathbf{F}\mathbf{e}_*) + \mathbf{e}_*' \boldsymbol{\Omega}^{-1} \mathbf{e}_*$$

Again, by generalized least-squares theory, we can show that

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{e}_*) = S(\boldsymbol{\phi}, \boldsymbol{\theta}) + (\mathbf{e}_* - \hat{\mathbf{e}}_*)' \mathbf{D} (\mathbf{e}_* - \hat{\mathbf{e}}_*)$$

where

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = S(\boldsymbol{\phi}, \boldsymbol{\theta}, \hat{\mathbf{e}}_*) = \hat{\mathbf{a}}' \hat{\mathbf{a}} + \hat{\mathbf{e}}_*' \boldsymbol{\Omega}^{-1} \hat{\mathbf{e}}_* \tag{A7.3.11}$$

is the unconditional sum-of-squares function and

$$\hat{\boldsymbol{e}}_* = E[\mathbf{e}_*|\mathbf{w}, \boldsymbol{\phi}, \boldsymbol{\theta}] = [\mathbf{e}_*] = \mathbf{D}^{-1} \mathbf{F}' \mathbf{L}_\theta'^{-1} \mathbf{L}_\theta^{-1} \mathbf{L}_\phi \mathbf{w} \tag{A7.3.12}$$

represents the conditional expectation of the preliminary values $\mathbf{e}_*$, with $\mathbf{D} = \boldsymbol{\Omega}^{-1} + \mathbf{F}' \mathbf{L}_\theta'^{-1} \mathbf{L}_\theta^{-1} \mathbf{F}$, and $\hat{\mathbf{a}} = [\mathbf{a}] = \mathbf{L}_\theta^{-1}(\mathbf{L}_\phi \mathbf{w} - \mathbf{F}\hat{\mathbf{e}}_*)$. By factorization of the joint distribution of $\mathbf{w}$ and $\mathbf{e}_*$, we can obtain

$$p(\mathbf{w}|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = (2\pi\sigma_a^2)^{-n/2} |\boldsymbol{\Omega}|^{-1/2} |\mathbf{D}|^{-1/2} \exp\left[-\frac{1}{2\sigma_a^2} S(\boldsymbol{\phi}, \boldsymbol{\theta})\right] \tag{A7.3.13}$$

as the unconditional likelihood. It follows immediately from (A7.3.13) that the maximum likelihood estimate for σ_a^2 is given by $\hat{\sigma}_a^2 = S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})/n$, where $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\theta}}$ denote maximum likelihood estimates.

Again, we note that $S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} [a_t]^2 + \hat{\mathbf{e}}_*' \boldsymbol{\Omega}^{-1} \hat{\mathbf{e}}_*$, and the elements $[a_1], [a_2], \ldots, [a_n]$ of $\hat{\mathbf{a}} = [\mathbf{a}]$ are computed recursively from the relation $\mathbf{L}_\theta[\mathbf{a}] = \mathbf{L}_\phi \mathbf{w} - \mathbf{F}[\mathbf{e}_*]$ as

$$[a_t] = w_t - \phi_1 [w_{t-1}] - \ldots - \phi_p [w_{t-p}] + \theta_1 [a_{t-1}] + \ldots + \theta_q [a_{t-q}]$$

for $t = 1, 2, \ldots, n$, using the back-forecasted values $[\mathbf{e}_*]$ for the preliminary values, with $[w_t] = w_t$ for $t = 1, 2, \ldots, n$. In addition, the back-forecasts $\hat{\mathbf{e}}_* = [\mathbf{e}_*]$ can be calculated from $\hat{\mathbf{e}}_* = \mathbf{D}^{-1} \mathbf{F}' \mathbf{u}$, where $\mathbf{u} = \mathbf{L}_\theta'^{-1} \mathbf{L}_\theta^{-1} \mathbf{L}_\phi \mathbf{w} = \mathbf{L}_\theta'^{-1} \mathbf{a}^0$, and the elements u_t of $\mathbf{u}$ are calculated through the backward recursion as

$$u_t = a_t^0 + \theta_1 u_{t+1} + \ldots + \theta_q u_{t+q}$$

with starting values $u_{n+1} = \ldots = u_{n+q} = 0$, and the a_t^0 are the elements of $\mathbf{a}^0 = \mathbf{L}_\theta^{-1} \mathbf{L}_\phi \mathbf{w}$ and denote the estimates of the a_t conditional on zero starting values $\mathbf{e}_* = \mathbf{0}$. Also, the

vector $\mathbf{h} = \mathbf{F}'\mathbf{u}$ consists of the $p + q$ elements:

$$h_j = \begin{cases} \sum_{i=1}^{j} \phi_{p-j+i} u_i & j = 1, \ldots, p \\ -\sum_{i=1}^{j-p} \theta_{q-j+p+i} u_i & j = p+1, \ldots, p+q \end{cases}$$

Finally, using (A7.1.1) and (A7.1.2), in $S(\boldsymbol{\phi}, \boldsymbol{\theta})$ we may write $\hat{\mathbf{e}}_*' \boldsymbol{\Omega}^{-1} \hat{\mathbf{e}}_* = \hat{\mathbf{a}}_*' \hat{\mathbf{a}}_* + (\hat{\mathbf{w}}_* - \mathbf{C}'\hat{\mathbf{a}}_*)' \mathbf{K}^{-1} (\hat{\mathbf{w}}_* - \mathbf{C}'\hat{\mathbf{a}}_*)$, so that we have

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1-q}^{n} [a_t]^2 + (\hat{\mathbf{w}}_* - \mathbf{C}'\hat{\mathbf{a}}_*)' \mathbf{K}^{-1} (\hat{\mathbf{w}}_* - \mathbf{C}'\hat{\mathbf{a}}_*) \tag{A7.3.14}$$

where $\mathbf{K} = \sigma_a^{-2} \Gamma_p - \mathbf{C}'\mathbf{C}$, as well as $|\boldsymbol{\Omega}| = |\mathbf{K}|$.

Therefore, in general, the likelihood associated with a series $\mathbf{z}$ of $n + d$ values generated by any ARIMA process is given by

$$L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2 | \mathbf{z}) = (2\pi\sigma_a^2)^{-n/2} \, |\mathbf{M}_n^{(p,q)}|^{1/2} \exp\left[-\frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_a^2} \right] \tag{A7.3.15}$$

where

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} [a_t]^2 + \hat{\mathbf{e}}_*' \boldsymbol{\Omega}^{-1} \hat{\mathbf{e}}_*$$

and $|\mathbf{M}_n^{(p,q)}| = |\boldsymbol{\Omega}|^{-1} |\mathbf{D}|^{-1} = |\mathbf{K}|^{-1} |\mathbf{D}|^{-1}$. Also, by expressing the mixed ARMA model as an infinite moving average $\tilde{w}_t = (1 + \psi_1 B + \psi_2 B^2 + \ldots) a_t$, and referring to results for the pure MA model, it follows that in the unconditional sum-of-squares function for the mixed model, we have the relation that $\hat{\mathbf{e}}_*' \boldsymbol{\Omega}^{-1} \hat{\mathbf{e}}_* = \sum_{t=-\infty}^{0} [a_t]^2$. Hence, we also have the representation $S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=-\infty}^{n} [a_t]^2$, and in practice the values $[a_t]$ may be computed recursively with the summation proceeding from some point $t = 1 - Q$, beyond which the $[a_t]$'s are negligible.

Special Case: AR(p). In the special case of a pure AR(p) model, the results described above simplify somewhat. We then have $\mathbf{e}_* = \mathbf{w}_*$, $\boldsymbol{\Omega} = \sigma_a^{-2} \boldsymbol{\Gamma}_p$, $\mathbf{L}_\theta = \mathbf{I}_n$, $\mathbf{D} = \sigma_a^2 \boldsymbol{\Gamma}_p^{-1} + \mathbf{F}'\mathbf{F} = \sigma_a^2 \boldsymbol{\Gamma}_p^{-1} + \mathbf{A}_p' \mathbf{A}_p$, and $\hat{\mathbf{w}}_* = \mathbf{D}^{-1} \mathbf{F}' \mathbf{L}_\phi \mathbf{w} = \mathbf{D}^{-1} \mathbf{A}_p' \mathbf{L}_{11} \mathbf{w}_p$, where $\mathbf{w}_p' = (w_1, w_2, \ldots, w_p)$ and $\mathbf{L}_{11}$ is the $p \times p$ upper left submatrix of $\mathbf{L}_\phi$. It can then be shown that the back-forecasts $\hat{w}_t$ are determined from the relations $\hat{w}_t = \phi_1 \hat{w}_{t+1} + \cdots + \phi_p \hat{w}_{t+p}$, $t = 0, -1, \ldots, 1 - p$, with $\hat{w}_t = w_t$ for $1 \le t \le n$, and hence these are the same as values obtained from the use of the backward model approach, as discussed in Section 7.1.4, for the special case of the AR model. Thus, we obtain the exact sum of squares as $S(\boldsymbol{\phi}) = \sum_{t=1}^{n} [a_t]^2 + \sigma_a^2 \hat{\mathbf{w}}_*' \boldsymbol{\Gamma}_p^{-1} \hat{\mathbf{w}}_*$.

To illustrate, consider the first-order autoregressive process in w_t,

$$w_t - \phi w_{t-1} = a_t \tag{A7.3.16}$$

where w_t might be the dth difference $\nabla^d z_t$ of the actual observations and a series $\mathbf{z}$ of length $n + d$ observations is available. To compute the likelihood (A7.3.15), we require

$$\begin{aligned} S(\phi) &= \sum_{t=1}^{n} [a_t]^2 + (1-\phi^2)\hat{w}_0^2 \\ &= \sum_{t=2}^{n} (w_t - \phi w_{t-1})^2 + (w_1 - \phi\hat{w}_0)^2 + (1-\phi^2)\hat{w}_0^2 \end{aligned}$$

since $\Gamma_1 = \gamma_0 = \sigma_a^2(1-\phi^2)^{-1}$. Now, because $\mathbf{D} = \sigma_a^2\Gamma_1^{-1} + \mathbf{A}_1'\mathbf{A}_1 = \sigma_a^2\gamma_0^{-1} + \phi^2 = 1$, and hence $\hat{w}_0 = \phi w_1$, substituting this into the last two terms of $S(\phi)$ above, it reduces to

$$S(\phi) = \sum_{t=2}^{n} (w_t - \phi w_{t-1})^2 + (1-\phi^2)w_1^2 \tag{A7.3.17}$$

as a result that may be obtained more directly by methods discussed in Appendix A7.4.

Special case: ARMA(1,1). As an example for the mixed model, consider the ARMA(1, 1) model

$$w_t - \phi w_{t-1} = a_t - \theta a_{t-1} \tag{A7.3.18}$$

Then, we have $\mathbf{e}_*' = (w_0, a_0)$, $\mathbf{A}_1 = \phi$, $B_1 = -\theta$, and

$$\sigma_a^2\mathbf{\Omega} = \sigma_a^2 \begin{bmatrix} \sigma_a^{-2}\gamma_0 & 1 \\ 1 & 1 \end{bmatrix}$$

with $\sigma_a^{-2}\gamma_0 = (1 + \theta^2 - 2\phi\theta)/(1-\phi^2)$. Thus, we have

$$\begin{aligned} \mathbf{D} = \mathbf{\Omega}^{-1} + \mathbf{F}'\mathbf{L}_\theta'^{-1}\mathbf{L}_\theta^{-1}\mathbf{F} &= \frac{1}{\sigma_a^{-2}\gamma_0 - 1} \begin{bmatrix} 1 & -1 \\ -1 & \sigma_a^{-2}\gamma_0 \end{bmatrix} \\ &+ \frac{1-\theta^{2n}}{1-\theta^2} \begin{bmatrix} \phi^2 & -\phi\theta \\ -\phi\theta & \theta^2 \end{bmatrix} \end{aligned}$$

and the estimates of the initial values are obtained as $\hat{\boldsymbol{e}}_* = \mathbf{D}^{-1}\mathbf{h}$, where $\mathbf{h}' = (h_1, h_2) = (\phi, -\theta)u_1$, the u_t are obtained from the backward recursion $u_t = a_t^0 + \theta u_{t+1}$, $u_{n+1} = 0$, and $a_t^0 = w_t - \phi w_{t-1}^0 + \theta a_{t-1}^0, t = 1, 2, \ldots, n$, are obtained using the zero initial values $w_0^0 = a_0^0 = 0$, with $w_t^0 = w_t$ for $1 \le t \le n$. Thus, the exact sum of squares is obtained as

$$S(\phi, \theta) = \sum_{t=0}^{n} [a_t]^2 + \frac{(\hat{w}_0 - \hat{a}_0)^2}{\sigma_a^{-2}\gamma_0 - 1} \tag{A7.3.19}$$

with $[a_t] = w_t - \phi[w_{t-1}] + \theta[a_{t-1}], t = 1, 2, \ldots, n$, and $\sigma_a^{-2}\gamma_0 - 1 = \mathbf{K} = (\phi - \theta)^2/(1 - \phi^2)$. In addition, we have $|\mathbf{M}_n^{(1,1)}| = \{|\mathbf{K}||\mathbf{D}|\}^{-1}$, with

$$|\mathbf{K}||\mathbf{D}| = 1 + \frac{1-\theta^{2n}}{1-\theta^2}\frac{(\phi-\theta)^2}{1-\phi^2}$$

APPENDIX A7.4 EXACT LIKELIHOOD FUNCTION FOR AN AUTOREGRESSIVE PROCESS

We now suppose that a given series $\mathbf{w}' = (w_1, w_2, \ldots, w_n)$ is generated by the pth-order stationary autoregressive model:

$$w_t - \phi_1 w_{t-1} - \phi_2 w_{t-2} - \cdots - \phi_p w_{t-p} = a_t$$

where, temporarily, the w_t's are assumed to have mean $\mu = 0$, but as before, the argument can be extended to the case where $\mu \neq 0$. Assuming normality for the a_t's and hence for the w_t's, the joint probability density function of the w_t's is

$$p(\mathbf{w}|\boldsymbol{\phi}, \sigma_a^2) = (2\pi\sigma_a^2)^{-n/2}|\mathbf{M}_n^{(p,0)}|^{1/2}\exp\left[-\frac{\mathbf{w}'\mathbf{M}_n^{(p,0)}\mathbf{w}}{2\sigma_a^2}\right] \tag{A7.4.1}$$

and because of the reversible character of the general process, the $n \times n$ matrix $\mathbf{M}_n^{(p,0)}$ is symmetric about *both* of its principal diagonals. Such a matrix is said to be *doubly* symmetric. Now,

$$p(\mathbf{w}|\boldsymbol{\phi}, \sigma_a^2) = p(w_{p+1}, w_{p+2}, \ldots, w_n|\mathbf{w}_p, \boldsymbol{\phi}, \sigma_a^2)p(\mathbf{w}_p, |\boldsymbol{\phi}, \sigma_a^2)$$

where $\mathbf{w}_p' = (w_1, w_2, \ldots, w_p)$. The first factor on the right may be obtained by making use of the distribution

$$p(a_{p+1}, \ldots, a_n) = (2\pi\sigma_a^2)^{-(n-p)/2}\exp\left[-\frac{1}{2\sigma_a^2}\sum_{t=p+1}^{n} a_t^2\right] \tag{A7.4.2a}$$

For fixed $\mathbf{w}_p$, $(a_{p+1}, \ldots, a_n)$ and $(w_{p+1}, \ldots, w_n)$ are related by the transformation

$$\begin{aligned} a_{p+1} &= w_{p+1} - \phi_1 w_p - \cdots - \phi_p w_1 \\ &\vdots \\ a_n &= w_n - \phi_1 w_{n-1} - \cdots - \phi_p w_{n-p} \end{aligned}$$

which has unit Jacobian. Thus, we obtain

$$\begin{aligned} &p(w_{p+1}, \ldots, w_n \mid \mathbf{w}_p, \phi, \sigma_a^2) \\ &\quad = (2\pi\sigma_a^2)^{-(n-p)/2}\exp\left[-\frac{1}{2\sigma_a^2}\sum_{t=p+1}^{n}(w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p})^2\right] \end{aligned} \tag{A7.4.2b}$$

Also,

$$p(\mathbf{w}_p, |\boldsymbol{\phi}, \sigma_a^2) = (2\pi\sigma_a^2)^{-p/2}|\mathbf{M}_p^{(p,0)}|^{1/2}\exp\left[-\frac{1}{2\sigma_a^2}\mathbf{w}_p'\mathbf{M}_p^{(p,0)}\mathbf{w}_p\right]$$

Thus,

$$p(\mathbf{w}|\boldsymbol{\phi}, \sigma_a^2) = (2\pi\sigma_a^2)^{-n/2}|\mathbf{M}_p^{(p,0)}|^{1/2}\exp\left[\frac{-S(\boldsymbol{\phi})}{2\sigma_a^2}\right] \tag{A7.4.3}$$

where

$$S(\boldsymbol{\phi}) = \sum_{i=1}^{p}\sum_{j=1}^{p} m_{ij}^{(p)} w_i w_j + \sum_{t=p+1}^{n} (w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p})^2 \tag{A7.4.4}$$

Also,

$$\mathbf{M}_p^{(p,0)} = \{m_{ij}^{(p)}\} = \{\gamma_{|i-j|}\}^{-1}\sigma_a^2 = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{bmatrix}^{-1} \sigma_a^2 \tag{A7.4.5}$$

where $\gamma_0, \gamma_1, \ldots, \gamma_{p-1}$ are the theoretical autocovariances of the process, and $|\mathbf{M}_p^{(p,0)}| = |\mathbf{M}_n^{(p,0)}|$.

Now, let $n = p + 1$, so that

$$\mathbf{w}'_{p+1}\mathbf{M}_{p+1}^{(p,0)}\mathbf{w}_{p+1} = \sum_{i=1}^{p}\sum_{j=1}^{p} m_{ij}^{(p)} w_i w_j + \left(w_{p+1} - \phi_1 w_p - \phi_2 w_{p-1} - \cdots - \phi_p w_1\right)^2$$

Then,

$$\mathbf{M}_{p+1}^{(p)} = \left[\begin{array}{ccc|c} & \mathbf{M}_p^{(p)} & & \begin{matrix}0\\0\\\vdots\\\vdots\end{matrix} \\ \hline 0 & 0 & \cdots & 0 \end{array}\right] + \left[\begin{array}{ccc|c} \phi_p^2 & \phi_p\phi_{p-1} & \cdots & -\phi_p \\ \phi_p\phi_{p-1} & \phi_{p-1}^2 & \cdots & -\phi_{p-1} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & -\phi_1 \\ \hline -\phi_p & -\phi_{p-1} & \cdots & 1 \end{array}\right]$$

and the elements of $\mathbf{M}_p^{(p)} = \mathbf{M}_p^{(p,0)}$ can now be deduced from the consideration that both $\mathbf{M}_p^{(p)}$ and $\mathbf{M}_{p+1}^{(p)}$ are doubly symmetric. Thus, for example,

$$\mathbf{M}_2^{(1)} = \begin{bmatrix} m_{11}^{(1)} + \phi_1^2 & -\phi_1 \\ -\phi_1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -\phi_1 \\ -\phi_1 & m_{11}^{(1)} + \phi_1^2 \end{bmatrix}$$

and after equating elements in the two matrices, we have

$$\mathbf{M}_1^{(1)} = m_{11}^{(1)} = 1 - \phi_1^2$$

Proceeding in this way, we find for processes of orders 1 and 2:

$$\mathbf{M}_1^{(1)} = 1 - \phi_1^2 \qquad |\mathbf{M}_1^{(1)}| = 1 - \phi_1^2$$

$$\mathbf{M}_2^{(2)} = \begin{bmatrix} 1 - \phi_2^2 & -\phi_1(1 + \phi_2) \\ -\phi_1(1 + \phi_2) & 1 - \phi_2^2 \end{bmatrix}$$

$$|\mathbf{M}_2^{(2)}| = (1 + \phi_2)^2[(1 - \phi_2)^2 - \phi_1^2]$$

For example, when $p = 1$,

$$p(\mathbf{w}|\phi, \sigma_a^2) = (2\pi\sigma_a^2)^{-n/2}(1-\phi^2)^{1/2}\exp\left\{-\frac{1}{2\sigma_a^2}\left[(1-\phi^2)w_1^2 + \sum_{t=2}^{n}(w_t - \phi w_{t-1})^2\right]\right\}$$

which checks with the result obtained in (A7.3.17). The process of generation must lead to matrices $\mathbf{M}_p^{(p)}$, whose elements are *quadratic* in the ϕ's.

Thus, it is clear from (A7.4.4) that not only is $S(\boldsymbol{\phi}) = \mathbf{w}'\mathbf{M}_n^{(p)}\mathbf{w}$ a quadratic form in the w_t's, but it is also quadratic in the parameters $\boldsymbol{\phi}$. Writing $\boldsymbol{\phi}_u' = (1, \phi_1, \phi_2, \ldots, \phi_p)$, it is clearly true that for some $(p+1)\times(p+1)$ matrix $\mathbf{D}$ whose elements are quadratic functions of the w_t's,

$$\mathbf{w}'\mathbf{M}_n^{(p)}\mathbf{w} = \boldsymbol{\phi}_u'\mathbf{D}\boldsymbol{\phi}_u$$

Now, write

$$\mathbf{D} = \begin{bmatrix} D_{11} & -D_{12} & -D_{13} & \cdots & -D_{1,p+1} \\ -D_{12} & D_{22} & D_{23} & \cdots & D_{2,p+1} \\ \vdots & \vdots & \vdots & & \vdots \\ -D_{1,p+1} & D_{2,p+1} & D_{3,p+1} & \cdots & D_{p+1,p+1} \end{bmatrix} \tag{A7.4.6}$$

Inspection of (A7.4.4) shows that the elements D_{ij} are ''symmetric'' sums of squares and lagged products, defined by

$$D_{ij} = D_{ji} = w_i w_j + w_{i+1}w_{j+1} + \cdots + w_{n+1-j}w_{n+1-i} \tag{A7.4.7}$$

where the sum D_{ij} contains $n - (i-1) - (j-1)$ terms.

Finally, we can write the *exact* probability density, and hence the exact likelihood, as

$$p(\mathbf{w}|\boldsymbol{\phi}, \sigma_a^2) = L(\boldsymbol{\phi}, \sigma_a^2|\mathbf{w}) = (2\pi\sigma_a^2)^{-n/2}|\mathbf{M}_p^{(p)}|^{1/2}\exp\left[\frac{-S(\boldsymbol{\phi})}{2\sigma_a^2}\right] \tag{A7.4.8}$$

where

$$S(\boldsymbol{\phi}) = \mathbf{w}_p'\mathbf{M}_p^{(p)}\mathbf{w}_p + \sum_{t=p+1}^{n}(w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p})^2 = \boldsymbol{\phi}_u'\mathbf{D}\boldsymbol{\phi}_u \tag{A7.4.9}$$

and the log-likelihood is

$$l(\boldsymbol{\phi}, \sigma_a^2|\mathbf{w}) = -\frac{n}{2}\ln(\sigma_a^2) + \frac{1}{2}\ln|\mathbf{M}_p^{(p)}| - \frac{S(\boldsymbol{\phi})}{2\sigma_a^2} \tag{A7.4.10}$$

For example, when $p = 1$, we have

$$\begin{aligned} S(\phi) &= (1-\phi^2)w_1^2 + \sum_{t=2}^{n}(w_t - \phi w_{t-1})^2 \\ &= \sum_{t=1}^{n} w_t^2 - 2\phi\sum_{t=2}^{n} w_{t-1}w_t + \phi^2\sum_{t=2}^{n-1} w_t^2 \equiv D_{11} - 2\phi D_{12} + \phi^2 D_{22} \end{aligned}$$

Maximum Likelihood Estimates. Differentiating with respect to σ_a^2 and each of the ϕ's in (A7.4.10), we obtain

$$\frac{\partial l}{\partial \sigma_a^2} = -\frac{n}{2\sigma_a^2} + \frac{S(\boldsymbol{\phi})}{2(\sigma_a^2)^2} \tag{A7.4.11}$$

$$\frac{\partial l}{\partial \phi_j} = M_j + \sigma_a^{-2}(D_{1,j+1} - \phi_1 D_{2,j+1} - \cdots - \phi_p D_{p+1,j+1})$$

$$j = 1, 2, \ldots, p \tag{A7.4.12}$$

where

$$M_j = \frac{\partial(\frac{1}{2}\ln|\mathbf{M}_p^{(p)}|)}{\partial \phi_j}$$

Hence, maximum likelihood estimates may be obtained by equating these expressions to zero and solving the resultant equations.

We have at once from (A7.4.11)

$$\hat{\sigma}_a^2 = \frac{S(\hat{\boldsymbol{\phi}})}{n} \tag{A7.4.13}$$

Estimates of $\boldsymbol{\phi}$. A difficulty occurs in dealing with equation (A7.4.12) since, in general, the quantities M_j $(j = 1, 2, \ldots, p)$ are complicated functions of the ϕ's. We consider briefly four alternative approximations.

1. *Least-Squares Estimates*. Since the expected value of $S(\boldsymbol{\phi})$ is proportional to n, while the value of $|\mathbf{M}_p^{(p)}|$ is independent of n, (A7.4.8) is for moderate or large sample sizes dominated by the term in $S(\boldsymbol{\phi})$ and the term in $|\mathbf{M}_p^{(p)}|$ is, by comparison, small.

If we ignore the influence of this term, then

$$l(\boldsymbol{\phi}, \sigma_a^2|\mathbf{w}) \simeq -\frac{n}{2}\ln(\sigma_a^2) - \frac{S(\boldsymbol{\phi})}{2\sigma_a^2} \tag{A7.4.14}$$

and the estimates $\hat{\boldsymbol{\phi}}$ of $\boldsymbol{\phi}$ obtained by maximization of (A7.4.14) are the least-squares estimates obtained by minimizing $S(\boldsymbol{\phi})$. Now, from (A7.4.9), $S(\boldsymbol{\phi}) = \boldsymbol{\phi}_u' \mathbf{D} \boldsymbol{\phi}_u$, where $\mathbf{D}$ is a $(p+1) \times (p+1)$ matrix of symmetric sums of squares and products, defined in (A7.4.7). Thus, on differentiating, the minimizing values are

$$\begin{aligned}
D_{12} &= \hat{\phi}_1 D_{22} + \hat{\phi}_2 D_{23} + \cdots + \hat{\phi}_p D_{2,p+1} \\
D_{13} &= \hat{\phi}_1 D_{23} + \hat{\phi}_2 D_{33} + \cdots + \hat{\phi}_p D_{3,p+1} \\
&\vdots \\
D_{1,p+1} &= \hat{\phi}_1 D_{2,p+1} + \hat{\phi}_2 D_{3,p+1} + \cdots + \hat{\phi}_p D_{p+1,p+1}
\end{aligned} \tag{A7.4.15}$$

which, in an obvious matrix notation, can be written as

$$\mathbf{d} = \mathbf{D}_p \hat{\boldsymbol{\phi}}$$

so that

$$\hat{\boldsymbol{\phi}} = \mathbf{D}_p^{-1}\mathbf{d}$$

These least-squares estimates also maximize the posterior density (7.5.15).

2. *Approximate Maximum Likelihood Estimates*. We now recall an earlier result (3.2.3), which may be written as

$$\gamma_j - \phi_1\gamma_{j-1} - \phi_2\gamma_{j-2} - \cdots - \phi_p\gamma_{j-p} = 0 \qquad j > 0 \tag{A7.4.16}$$

Also, on taking expectations in (A7.4.12) and using the fact that $E[\partial l/\partial\phi_j] = 0$, we obtain

$$M_j\sigma_a^2 + (n-j)\gamma_j - (n-j-1)\phi_1\gamma_{j-1} - (n-j-2)\phi_2\gamma_{j-2} - \cdots - (n-j-p)\phi_p\gamma_{j-p} = 0 \tag{A7.4.17}$$

After multiplying (A7.4.16) by n and subtracting the result from (A7.4.17), we obtain

$$M_j\sigma_a^2 = j\gamma_j - (j+1)\phi_1\gamma_{j-1} - \cdots - (j+p)\phi_p\gamma_{j-p}$$

Therefore, on using $D_{i+1,j+1}/(n-j-i)$ as an estimate of $\gamma_{|j-i|}$, a natural estimate of $M_j\sigma_a^2$ is

$$j\frac{D_{1,j+1}}{n-j} - (j+1)\phi_1\frac{D_{2,j+1}}{n-j-1} - \cdots - (j+p)\phi_p\frac{D_{p+1,j+1}}{n-j-p}$$

Substituting this estimate in (A7.4.12) yields

$$\frac{\partial l}{\partial\phi_j} \simeq n\sigma_a^{-2}\left(\frac{D_{1,j+1}}{n-j} - \phi_1\frac{D_{2,j+1}}{n-j-1} - \cdots - \phi_p\frac{D_{p+1,j+1}}{n-j-p}\right) \qquad j = 1, 2, \ldots, p \tag{A7.4.18}$$

leading to a set of linear equations of the form (A7.4.15), but now with

$$D_{ij}^* = \frac{nD_{ij}}{n-(i-1)-(j-1)}$$

replacing D_{ij}.

3. *Conditional Least-Squares Estimates*. For moderate and relatively large n, we might also consider the conditional sum-of-squares function, obtained by adopting the procedure in Section 7.1.3. This yields the sum of squares given in the exponent of the expression in (A7.4.2),

$$S_*(\boldsymbol{\phi}) = \sum_{t=p+1}^{n}(w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p})^2$$

and is the sum of squares associated with the conditional distribution of $w_{p+1}, \ldots, w_n$, given $\mathbf{w}_p' = (w_1, w_2, \ldots, w_p)$. Conditional least-squares estimates are obtained by minimizing $S_*(\boldsymbol{\phi})$, which is a standard linear least-squares regression problem

associated with the linear model $w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \cdots + \phi_p w_{t-p} + a_t, t = p+1, \ldots, n$. This results in the familiar least-squares estimates $\hat{\boldsymbol{\phi}} = \tilde{\mathbf{D}}_p^{-1}\tilde{\mathbf{d}}$, as in (A7.2.5), where $\tilde{\mathbf{D}}_p$ has (i, j) th element $\tilde{D}_{ij} = \sum_{t=p+1}^{n} w_{t-i} w_{t-j}$ and $\tilde{\mathbf{d}}$ has ith element $\tilde{d}_i = \sum_{t=p+1}^{n} w_{t-i} w_t$.

4. *Yule–Walker Estimates*. Finally, if n is moderate or large, as an approximation, we may replace the symmetric sums of squares and products in (A7.4.15) by n times the appropriate autocovariance estimate. For example, D_{ij}, where $|i - j| = k$, would be replaced by $nc_k = \sum_{t=1}^{n-k} \tilde{w}_t \tilde{w}_{t+k}$. On dividing by nc_0 throughout in the resultant equations, we obtain the following relations expressed in terms of the estimated autocorrelations $r_k = c_k/c_0$:

$$\begin{aligned} r_1 &= \hat{\phi}_1 + \hat{\phi}_2 r_1 + \cdots + \hat{\phi}_p r_{p-1} \\ r_2 &= \hat{\phi}_1 r_1 + \hat{\phi}_2 + \cdots + \hat{\phi}_p r_{p-2} \\ &\vdots \\ r_p &= \hat{\phi}_1 r_{p-1} + \hat{\phi}_2 r_{p-2} + \cdots + \hat{\phi}_p \end{aligned}$$

These are the well-known Yule–Walker equations.

In the matrix notation (7.3.1), they can be written $\mathbf{r} = \mathbf{R}\hat{\boldsymbol{\phi}}$, so that

$$\hat{\boldsymbol{\phi}} = \mathbf{R}^{-1}\mathbf{r} \tag{A7.4.19}$$

which corresponds to equations (3.2.7), with $\mathbf{r}$ substituted for $\boldsymbol{\rho}_p$ and $\mathbf{R}$ for $\mathbf{P}_p$.

To illustrate the differences among the four estimates, take the case $p = 1$. Then, $M_1\sigma_a^2 = -\gamma_1$ and, corresponding to (A7.4.12), the exact maximum likelihood estimate of ϕ is the solution of

$$-\gamma_1 + D_{12} - \phi D_{22} \equiv -\gamma_1 + \sum_{t=2}^{n} w_t w_{t-1} - \phi \sum_{t=2}^{n-1} w_t^2 = 0$$

Note that $\gamma_1 = \sigma_a^2 \phi/(1 - \phi^2)$ and the maximum likelihood solution for $\sigma_a^2, \hat{\sigma}_a^2 = S(\phi)/n$ from (A7.4.13), can be substituted in the expression for γ_1 in the likelihood equation above, where $S(\phi) = D_{11} - 2\phi D_{12} + \phi^2 D_{22}$ as in (A7.4.9). This results in a *cubic* equation in ϕ, whose solution yields the maximum likelihood estimate of ϕ. Upon rearranging, the cubic equation for $\hat{\phi}$ can be written as

$$(n-1)D_{22}\hat{\phi}^3 - (n-2)D_{12}\hat{\phi}^2 - (nD_{22} + D_{11})\hat{\phi} + nD_{12} = 0 \tag{A7.4.20}$$

and there is a single *unique* solution to this cubic equation such that $-1 < \hat{\phi} < 1$ (e.g., Anderson, 1971, p. 354).

Approximation 1 corresponds to ignoring the term γ_1 altogether, yielding

$$\hat{\phi} = \frac{\sum_{t=2}^{n} w_t w_{t-1}}{\sum_{t=2}^{n-1} w_1^2} = \frac{D_{12}}{D_{22}}$$

Approximation 2 corresponds to substituting the estimate $\sum_{t=2}^{n} w_t w_{t-1}/(n-1)$ for γ_1, yielding

$$\hat{\phi} = \frac{\sum_{t=2}^{n} w_t w_{t-1}/(n-1)}{\sum_{t=2}^{n-1} w_t^2/(n-2)} = \frac{n-2}{n-1}\frac{D_{12}}{D_{22}}$$

Approximation 3 corresponds to the standard linear model least-squares estimate obtained by regression of w_t on w_{t-1} for $t = 2, 3, \ldots, n$, so that

$$\hat{\phi} = \frac{\sum_{t=2}^{n} w_t w_{t-1}}{\sum_{t=2}^{n} w_{t-1}^2} = \frac{D_{12}}{D_{22} + w_1^2}$$

In effect, this can be viewed as obtained by substituting ϕw_1^2 for γ_1 in the likelihood equation above for ϕ.

Approximation 4 replaces the numerator and denominator by standard autocovariance estimates (2.1.12), yielding

$$\hat{\phi} = \frac{\sum_{t=2}^{n} w_t w_{t-1}}{\sum_{t=1}^{n} w_1^2} = \frac{c_1}{c_0} = r_1 = \frac{D_{12}}{D_{11}}$$

Usually, as in this example, for moderate and large samples, the differences between the estimates given by the various approximations will be small. We have often employed the least-squares estimates given by approximation 1 which can be computed directly from (A7.4.15). However, for computer calculations, it is often simplest, even when the fitted model is autoregressive, to use the general iterative algorithm described in Section 7.2.1, which computes least-squares estimates for any ARMA process.

Estimate of σ_a^2. Using approximation 4 with (A7.4.9) and (A7.4.13),

$$\hat{\sigma}_a^2 = \frac{S(\hat{\boldsymbol{\phi}})}{n} = c_0 \left[1 : \hat{\boldsymbol{\phi}}'\right] \left[\begin{array}{c|c} 1 & -\mathbf{r}' \\ \hline -\mathbf{r} & \mathbf{R} \end{array}\right] \left[\begin{array}{c} 1 \\ \hline \hat{\boldsymbol{\phi}} \end{array}\right]$$

On multiplying out the right-hand side and recalling that $\mathbf{r} - \mathbf{R}\hat{\boldsymbol{\phi}} = \mathbf{0}$, we find that

$$\hat{\sigma}_a^2 = c_0(1 - \mathbf{r}'\hat{\boldsymbol{\phi}}) = c_0(1 - \mathbf{r}'\mathbf{R}^{-1}\mathbf{r}) = c_0(1 - \hat{\boldsymbol{\phi}}'\mathbf{R}\hat{\boldsymbol{\phi}}) \tag{A7.4.21a}$$

It is readily shown that σ_a^2 can be similarly written in terms of the *theoretical* autocorrelations:

$$\sigma_a^2 = \gamma_0(1 - \boldsymbol{\rho}'\boldsymbol{\phi}) = \gamma_0(1 - \boldsymbol{\rho}'\mathbf{P}_p^{-1}\boldsymbol{\rho}) = \gamma_0(1 - \boldsymbol{\phi}'\mathbf{P}_p\boldsymbol{\phi}) \tag{A7.4.21b}$$

agreeing with the result (3.2.8).

Parallel expressions for $\hat{\sigma}_a^2$ may be obtained for approximations 1, 2, and 3.

Information Matrix. Differentiating for a second time in (A7.4.11) and (A7.4.18), we obtain

$$-\frac{\partial^2 l}{\partial(\sigma_a^2)^2} = -\frac{n}{2(\sigma_a^2)^2} + \frac{S(\boldsymbol{\phi})}{(\sigma_a^2)^3} \tag{A7.4.22a}$$