to allow for several inputs $X_{1,t}, X_{2,t}, \ldots, X_{m,t}$. Thus,

$$Y_t = \upsilon_1(B)X_{1,t} + \cdots + \nu_m(B)X_{m,t} + N_t \tag{A12.1.9}$$

$$= \delta_1^{-1}(B)\omega_1(B)X_{1,t-b_1} + \cdots + \delta_m^{-1}(B)\omega_m(B)X_{m,t-b_m} + N_t \tag{A12.1.10}$$

where $\upsilon_j(B) = \delta_j^{-1}(B)\omega_j(B)B^{b_j}$ is the generating function of the impulse response weights relating $X_{j,t}$ to the output Y_t. We assume, as before, that after differencing, (A12.1.9) may be written as

$$y_t = \upsilon_1(B)x_{1,t} + \cdots + \upsilon_m(B)x_{m,t} + n_t$$

where $y_t, x_{1,t}, \ldots, x_{m,t}$, and n_t are all jointly stationary processes. Multiplying throughout by $x_{1,t-k}, x_{2,t-k}, \ldots, x_{m,t-k}$ in turn, taking expectations, and forming the generating functions, we obtain

$$\begin{aligned}
\gamma^{x_1y}(B) &= \upsilon_1(B)\gamma^{x_1x_1}(B) + \upsilon_2(B)\gamma^{x_1x_2}(B) + \cdots + \upsilon_m(B)\gamma^{x_1x_m}(B) \\
\gamma^{x_2y}(B) &= \upsilon_1(B)\gamma^{x_2x_1}(B) + \upsilon_2(B)\gamma^{x_2x_2}(B) + \cdots + \upsilon_m(B)\gamma^{x_2x_m}(B) \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
\gamma^{x_my}(B) &= \upsilon_1(B)\gamma^{x_mx_1}(B) + \upsilon_2(B)\gamma^{x_mx_2}(B) + \cdots + \upsilon_m(B)\gamma^{x_mx_m}(B)
\end{aligned} \tag{A12.1.11}$$

On substituting $B = e^{-i2\pi f}$, the spectral equations are obtained. For example, with $m = 2$,

$$p_{x_1y}(f) = H_1(f)p_{x_1x_1}(f) + H_2(f)p_{x_1x_2}(f)$$
$$p_{x_2y}(f) = H_1(f)p_{x_2x_1}(f) + H_2(f)p_{x_2x_2}(f)$$

and the frequency response functions $H_1(f) = \upsilon_1(e^{-i2\pi f})$ and $H_2(f) = \upsilon_2(e^{-i2\pi f})$ can be calculated as described in Jenkins and Watts (1968). The impulse response weights can then be obtained using the inverse transformation (A12.1.6).

APPENDIX A12.2 CHOICE OF INPUT TO PROVIDE OPTIMAL PARAMETER ESTIMATES

Suppose that the input to a dynamic system can be made to follow an imposed stochastic process that is our choice. For example, it might be an autoregressive process, a moving average process, or white noise. To illustrate the problems involved in the optimal selection of this stochastic process, it is sufficient to consider an elementary example.

A12.2.1 Design of Optimal Inputs for a Simple System

Suppose that a system is under study for which the transfer function–noise model is assumed to be

$$Y_t = \beta_1 Y_{t-1} + \beta_2 X_{t-1} + a_t \qquad |\beta_1| < 1 \tag{A12.2.1}$$

where a_t is white noise. It is also assumed that the input and output processes are stationary and that X_t and Y_t denote deviations of these processes from their respective means. For large samples, and associated with any fixed probability, the approximate area of the Bayesian HPD region for β_1 and β_2, and also of the corresponding confidence region, is

proportional to $\Delta^{-1/2}$, where Δ is the determinant

$$\Delta = \begin{vmatrix} E[Y_t^2] & E[Y_t X_t] \\ E[Y_t X_t] & E[X_t^2] \end{vmatrix}$$

We will proceed by attempting to find the design minimizing the area of the HPD or confidence region and thus maximizing Δ. Now

$$\begin{aligned} E[Y_t^2] &= \sigma_Y^2 = \sigma_X^2 \beta_2^2 \frac{1+2q}{1-\beta_1^2} + \frac{\sigma_a^2}{1-\beta_1^2} \\ E[Y_t X_t] &= \sigma_X^2 \frac{\beta_2}{\beta_1} q \\ E[X_t^2] &= \sigma_X^2 \end{aligned} \tag{A12.2.2}$$

where

$$q = \sum_{i=1}^{\infty} \beta_1^i \rho_i \qquad \sigma_X^2 \rho_i = E[X_t X_{t-i}]$$

The value of the determinant may be written in terms of σ_X^2 as

$$\Delta = \frac{\sigma_X^2 \sigma_a^2}{1-\beta_1^2} + \frac{\beta_2^2 \sigma_X^4}{(1-\beta_1^2)^2} - \frac{\sigma_X^4 \beta_2^2}{\beta_1^2} \left(q - \frac{\beta_1^2}{1-\beta_1^2} \right)^2 \tag{A12.2.3}$$

Thus, as might be expected, the area of the region can be made small by making σ_X^2 large (i.e., by varying the input variable over a wide range). In practice, there may be limits to the amount of variation that can be allowed in X. Let us proceed by first supposing that σ_X^2 is held fixed at some specified value.

Solution with σ_X^2 Fixed. With $(1-\beta_1^2) > 0$ and for any fixed σ_X^2, we see from (A12.2.3) that Δ is maximized by setting

$$q = \frac{\beta_1^2}{1-\beta_1^2}$$

that is,

$$\beta_1 \rho_1 + \beta_1^2 \rho_2 + \beta_1^3 \rho_3 + \cdots = \beta_1^2 + \beta_1^4 + \beta_1^6 + \cdots$$

There are an infinite number of ways in which, for given β_1, this equality could be achieved. One obvious solution is

$$\rho_i = \beta_1^i$$

Thus, one way to maximize Δ for fixed σ_X^2 would be to force the input to follow the autoregressive process

$$(1 - \beta_1 B)X_t = \alpha_t$$

where α_t is a white noise process with variance $\sigma_\alpha^2 = \sigma_X^2(1 - \beta_1^2)$.

Solution with σ_Y^2 Fixed. So far we have supposed that σ_Y^2 is unrestricted. In some cases, we might wish to avoid too great a variation in the output rather than in the input. Suppose that σ_Y^2 is held equal to some fixed acceptable value but that σ_X^2 is unrestricted. Then the value of the determinant Δ can be written in terms of σ_Y^2 as

$$\Delta = \frac{\sigma_Y^4}{\beta_2^2}\left[\frac{\sigma_Y^2 - \sigma_a^2}{\sigma_Y^2} - \frac{\beta_1^2}{s^2}\left(\frac{q+s}{1+2q}\right)^2\right] \tag{A12.2.4}$$

where

$$s = \frac{\beta_1^2 r}{1 + \beta_1^2 r} \tag{A12.2.5}$$

and

$$r = \frac{\sigma_Y^2}{\sigma_Y^2 - \sigma_a^2} \tag{A12.2.6}$$

The maximum is achieved by setting

$$q = -s = \frac{-\beta_1^2 r}{1 + \beta_1^2 r} \tag{A12.2.7}$$

that is,

$$\beta_1\rho_1 + \beta_1^2\rho_2 + \beta_1^3\rho_3 + \cdots = -\beta_1^2 r + \beta_1^4 r^2 - \beta_1^6 r^3 + \cdots$$

There are again infinite ways of satisfying this equality. In particular, one solution is

$$\rho_i = (-\beta_1 r)^i \tag{A12.2.8}$$

which can be obtained by forcing the input to follow the autoregressive process

$$(1 + \beta_1 r B)X_t = \alpha_t \tag{A12.2.9}$$

where α_t is a white noise process with variance $\sigma_\alpha^2 = \sigma_X^2(1 - \beta_1^2 r^2)$. Since r is essentially positive, the sign of the parameter $(-\beta_1 r)$ of this autoregressive process is opposite to that obtained for the optimal input with σ_X^2 fixed.

Solution with $\sigma_Y^2 \times \sigma_X^2$ Fixed. In practice, it might happen that excessive variations in input and output were both to be avoided. If it were true that a given *percentage* decrease in the variance of X was equally as desirable as the same *percentage* decrease in the variance of Y, it would be sensible to maximize Δ subject to a fixed value of the product $\sigma_X^2 \times \sigma_Y^2$.

The determinant is

$$\Delta = \sigma_X^2 \sigma_Y^2 - \frac{\sigma_X^4 \beta_2^2 q^2}{\beta_1^2} \tag{A12.2.10}$$

which is maximized for fixed $\sigma_X^2 \sigma_Y^2$ only if $q = 0$. Once again there are an infinite number of solutions. However, by using a white noise input, Δ is maximized *whatever the value of* β_1. For such an input, using (A12.2.2), σ_X^2 is the positive root of

$$\sigma_X^4 \beta_2^2 + \sigma_X^2 \sigma_a^2 - k(1 - \beta_1^2) = 0 \tag{A12.2.11}$$

where $k = \sigma_X^2 \sigma_Y^2$, which is fixed.

A12.2.2 Numerical Example

Suppose that we were studying the first-order dynamic system (A12.2.1) with $\beta_1 = 0.50$ and $\beta_2 = 1.00$, so that

$$Y_t = 0.50 Y_{t-1} + 1.00 X_{t-1} + \alpha_t$$

where $\sigma_\alpha^2 = 0.2$.

σ_X^2 Fixed, σ_Y^2 Unrestricted. Suppose at first that the design is chosen to maximize Δ with $\sigma_X^2 = 1.0$. Then one optimal choice for the input X_t will be the autoregressive process

$$(1 - 0.5B) X_t = \alpha_t$$

where the white noise process α_t would have variance $\sigma_\alpha^2 = \sigma_X^2(1 - \beta_1^2) = 0.75$. Using (A12.2.2), the variance σ_Y^2 of the output would be 2.49, and the scheme will achieve a Bayesian region for β_1 and β_2 whose area is proportional to $\Delta^{-1/2} = 0.70$.

σ_Y^2 Fixed, σ_X^2 Unrestricted. The above scheme is optimal under the assumption that the input variance is $\sigma_X^2 = 1$ and the output variance is unrestricted. This output variance then turns out to be $\sigma_Y^2 = 2.49$. If, instead, the input variance were unrestricted, then with a *fixed* output variance of 2.49, we could, of course, do considerably better. In fact, using (A12.2.6), $r = 1.087$ and hence $\beta_1 r \simeq 0.54$, so that from (A12.2.9) one optimal choice for the unrestricted input would be the autoregressive process

$$(1 + 0.54B) X_t = \alpha_t$$

where in this case α_t is a white noise process with $\sigma_\alpha^2 = \sigma_X^2(1 - \beta_1^2 r^2)$. Using (A12.2.2) with $\sigma_Y^2 = 2.49$ fixed and $q = -0.214$ from (A12.2.7), the variance σ_X^2 of the input would now be increased to 2.91, so that $\sigma_\alpha^2 = 2.05$, and $\Delta^{-1/2}$, which measures the area of the Bayesian region, would be reduced to $\Delta^{-1/2} = 0.42$.

Product $\sigma_Y^2 \times \sigma_X^2$ Fixed. Finally, we consider a scheme that attempts to control both σ_Y^2 and σ_X^2 by maximizing Δ with $\sigma_Y^2 \times \sigma_X^2$ fixed. In the previous example in which σ_Y^2 was fixed, we found that $\Delta^{-1/2} = 0.42$ with $\sigma_X^2 = 2.91$ and $\sigma_Y^2 = 2.49$, so that the product is

$2.91 \times 2.49 = 7.25$. If our objective had been to minimize $\Delta^{-1/2}$ while keeping this product equal to 7.25, we could have made an optimal choice *without knowledge of* β_1 by choosing a white noise input $X_t = \alpha_t$. Using (A12.2.11), $\sigma_X^2 = \sigma_\alpha^2 = 2.29$, $\sigma_Y^2 = 3.16$, and in this case, as expected, $\Delta^{-1/2} = 0.37$, slightly smaller than that in the previous example.

It is worth considering this example in terms of spectral ideas. To optimize with σ_X^2 fixed, we have used an autoregressive input with ϕ_x positive that has high power at low frequencies. Since the gain of the system is high at low frequencies, this achieves maximum transfer from X to Y and so induces large variations in Y. When σ_Y^2 is fixed, we have introduced an input that is an autoregressive process with ϕ_x negative. This has high power at high frequencies. Since there is minimum transfer from X to Y at high frequencies, the disturbance in X must now be made large at these frequencies. When the product $\sigma_X^2 \times \sigma_Y^2$ is fixed, the ''compromise'' input white noise is indicated and does not require knowledge of β_1. This final maximization of Δ is equivalent to minimizing the (magnitude of the) correlation between the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, and in fact the correlation between these estimates is zero when a white noise input is used.

Conclusions. This investigation shows the following:

1. The optimal choice of design rests heavily on how we define ''optimal.''
2. Both in the case where α_X^2 is held fixed and in the case where α_Y^2 is held fixed, the optimal choices require specific stochastic processes for the input X_t whose parameters are functions of the *unknown* dynamic parameters. Thus, we are in the familiar paradoxical situation where we can do a better job of data gathering only to the extent that we already know something about the answer we seek. A sequential approach, where we improve the design as we find out more about the parameters, is a possibility worth further investigation. In particular, a pilot investigation using a possibly nonoptimal input, say white noise, could be used to generate data from which preliminary estimates of the dynamic parameters could be obtained. These estimates could then be used to specify a further input using one of our previous criteria.
3. The use of white noise is shown, *for the simple case investigated*, to be optimal for a sensible criterion of optimality, and its use as an input requires no prior knowledge of the parameters.

EXERCISES

12.1. Estimate of the cross-correlation function at lags $-1, 0$, and $+1$ for the following series of five pairs of observations:

t	1	2	3	4	5
x_t	11	7	8	12	14
y_t	7	10	6	7	10

12.2. If two series may be represented in ψ-weight form as

$$y_t = \psi_y(B)a_t \qquad x_t = \psi_x(B)a_t$$

(a) Show that their cross-covariance generating function

$$\gamma^{xy}(B) = \sum_{k=-\infty}^{\infty} \gamma_{xy}(k) B^k$$

is given by $\sigma_a^2 \psi_y(B)\psi_x(F)$.

(b) Use the above result to obtain the cross-covariance function between y_t and x_t when

$$y_t = (1 - \theta B)a_t \qquad x_t = (1 - \theta_1' B - \theta_2' B^2)a_t$$

12.3. After estimating a prewhitening transformation $\theta_x^{-1}(B)\phi_x(B)x_t = \alpha_t$ for an input series x_t and then computing the transformed output $\beta_t = \theta_x^{-1}(B)\phi_x(B)y_t$, cross-correlations $r_{\alpha\beta}(k)$ were obtained as follows:

k	$r_{\alpha\beta}(k)$	k	$r_{\alpha\beta}(k)$
0	0.05	5	0.24
1	0.31	6	0.07
2	0.52	7	−0.03
3	0.43	8	0.10
4	0.29	9	0.07

with $\hat{\sigma}_\alpha = 1.26$, $\hat{\sigma}_\beta = 2.73$, and $n = 187$.

(a) Obtain approximate standard errors for the cross-correlations.

(b) Calculate rough estimates for the impulse response weights v_j of a transfer function between y_t and x_t.

(c) Suggest a model form for the transfer function and give rough estimates of its parameters.

12.4. It is frequently the case that the user of an estimated transfer function–noise model $y_t = \delta^{-1}(B)\omega(B)B^b x_t + n_t$ will want to establish whether the steady-state gain $g = \delta^{-1}(1)\omega(1)$ makes good sense.

(a) For the first-order transfer function system

$$y_t = \frac{\omega_0}{1 - \delta B} x_{t-1}$$

show that an approximate standard error $\hat{\sigma}(\hat{g})$ of the estimate $\hat{g} = \hat{\omega}_0/(1 - \hat{\delta})$ is given by

$$\frac{\hat{\sigma}^2(\hat{g})}{\hat{g}^2} \simeq \frac{\text{var}[\hat{\omega}_0]}{\hat{\omega}_0^2} + \frac{\text{var}[\hat{\delta}]}{(1 - \hat{\delta})^2} + \frac{2\text{cov}[\hat{\omega}_0, \hat{\delta}]}{\hat{\omega}_0(1 - \hat{\delta})}$$

(b) Calculate $\hat{g}$ and an approximate value for $\hat{\sigma}(\hat{g})$ when $\hat{\omega}_0 = 5.2$, $\hat{\delta} = 0.65$, $\hat{\sigma}(\hat{\omega}_0) = 0.5$, $\hat{\sigma}(\hat{\delta}) = 0.1$, and $\text{cov}[\hat{\omega}_0, \hat{\delta}] = 0.025$.

12.5. Consider the regression model

$$Y_t = \beta_1 X_{1,t} + \beta_2 X_{2,t} + N_t$$

where N_t is a nonstationary error term following an IMA(0, 1, 1) process $\nabla N_t = a_t - \theta a_{t-1}$. Show that the regression model may be rewritten in the form

$$Y_t - \bar{Y}_{t-1} = \beta_1(X_{1,t} - \bar{X}_{1,t-1}) + \beta_2(X_{2,t} - \bar{X}_{2,t-1}) + a_t$$

where $\bar{Y}_{t-1}$, $\bar{X}_{1,t-1}$, and $\bar{X}_{2,t-1}$ are exponentially weighted moving averages so that, for example,

$$\bar{Y}_{t-1} = (1-\theta)(Y_{t-1} + \theta Y_{t-2} + \theta^2 Y_{t-3} + \cdots)$$

It will be seen that the fitting of this regression model with nonstationary noise by maximum likelihood is equivalent to fitting the *deviations* of the independent and dependent variables from *local updated exponentially weighted moving averages* by ordinary least-squares. (Refer to Section 9.5.1 for related ideas regarding transformation of regression models with autocorrelated noise N_t.)

12.6. Quarterly measurements of unemployment and 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; see also http://pages.stat.wisc.edu/ reinsel/bjr-data/.

(a) Plot the two time series using R.

(b) Calculate and plot the autocorrelation and partial autocorrelation functions of the two series. Repeat the calculations for the first differences of the two series. Would a variance stabilizing transformation be helpful for model development?

(c) Calculate and plot the cross-correlation function between the two series.

12.7. Refer to Exercise 12.6. Build (identify, estimate, and check) a transfer function–noise model that uses the GDP series X_t as input to help explain variations in the logged unemployment series Y_t.

12.8. Consider the transfer function–noise model fitted to the gas furnace data in (12.4.1) and (12.4.2). Note that the estimate of δ_2 is very close to zero. Re-estimate the parameters of this model setting δ_2 equal to zero. Describe the resulting impact on the estimate of the residual variance and other model parameters.

12.9. A bivariate time series consisting of sales data and a leading indicator is listed as Series M in Part Five of this book. The series is also available as ''BJsales'' in the datasets package of R.

(a) Plot the two time series using R.

(b) Calculate and plot the autocorrelation and partial autocorrelation functions of the two series. Find a suitable model for the leading indicator series.

(c) Calculate and plot the cross-correlation function between the two variables.

(d) Calculate and plot the cross-correlation function after prewhitening the series using the time series model developed in part (b).

(e) Estimate the impulse response function v_k for the two series.

12.10. Refer to Exercise 12.9. A bivariate transfer function–noise model was given for these series in Section 12.5.3.

(a) Use the results from Exercise 12.9 to justify the choice of transfer function model. Derive preliminary estimates of the parameters in this model.

(b) Justify the choice of the noise model given in Section 12.5.3.

(c) Estimate the parameters of the combined transfer function–noise model and perform the appropriate diagnostic checks on the fitted model.

13

INTERVENTION ANALYSIS, OUTLIER DETECTION, AND MISSING VALUES

Time series are often affected by special events or conditions such as policy changes, strikes, advertising promotions, environmental regulations, and similar events, which we will refer to as *intervention* events. In Section 13.1, we describe the method of intervention analysis, which can account for the expected effects of these interventions. For this, the transfer function models of the previous chapters are used, but in the intervention analysis model, the input series will be in the form of a simple pulse or step indicator function to signal the presence or absence of the event. The timing of the intervention event is assumed to be known in this analysis. Section 13.2 considers the related problem of detecting outlying or unusual behavior in a time series at an unknown point of time. Depending on how the outlier enters and its likely impact on the time series, two types of outlier models, additive outlier (AO) and innovational outlier (IO) models, are considered. A somewhat related problem of missing values in a time series is discussed in Section 13.3. The key focus of this section is on parameter estimation and evaluation of the likelihood function of an ARMA model for time series with missing values. However, consideration is also given to estimation of the missing values in the series.

13.1 INTERVENTION ANALYSIS METHODS

13.1.1 Models for Intervention Analysis

In the setting of intervention analysis, it is assumed that an intervention event has occurred at a known point in time T of a time series. It is of interest to determine whether there is

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

any evidence of a change or effect, of an expected kind, on the time series Y_t associated with the event. We consider the use of transfer function models to model the nature of and estimate the magnitude of the effects of the intervention, and hence to account for the possible unusual behavior in the time series related to the event. Based on the study by Box and Tiao (1975), the type of model we consider has the form

$$Y_t = \frac{\omega(B)B^b}{\delta(B)}\xi_t + N_t \tag{13.1.1}$$

where the term $\mathcal{Y}_t = \delta^{-1}(B)\omega(B)B^b\xi_t$ represents the effects of the intervention event in terms of the deterministic input series ξ_t, and N_t is the noise series that represents the underlying time series without the intervention effects. It is assumed that N_t follows an ARIMA(p, d, q) model, $\varphi(B)N_t = \theta(B)a_t$, with $\varphi(B) = \phi(B)(1 - B)^d$. Multiplicative seasonal ARIMA models as presented in Chapter 9 can also be included for N_t, but special note of the seasonal models will not be made in this chapter.

There are two common types of deterministic input variables ξ_t that have been found useful to represent the impact of intervention events on a time series. Both of these are indicator variables taking only the values 0 and 1 to denote the nonoccurrence and occurrence of the intervention. One type is a *step function* at time T, given by

$$S_t^{(T)} = \begin{cases} 0 & t < T \\ 1 & t \geq T \end{cases} \tag{13.1.2}$$

which would typically be used to represent the effects of an intervention that are expected to remain permanently after time T to some extent. The other type is a *pulse function* at T, given by

$$P_t^{(T)} = \begin{cases} 0 & t \neq T \\ 1 & t = T \end{cases} \tag{13.1.3}$$

which could represent the effects of an intervention that are temporary or transient and will die out after time T. These indicator input variables are used in many situations where the effects of the intervention cannot be represented as the response to a quantitative variable because such a quantitative variable does not exist or it is impractical or impossible to obtain measurements on such a variable.

Because of the deterministic nature of the indicator input series ξ_t in (13.1.1), unlike the transfer function model situation of Chapter 12, identification of the structure of the intervention model operator $v(B) = \delta^{-1}(B)\omega(B)B^b$ cannot be based on the technique of prewhitening. Instead, it is necessary to postulate the form of the intervention model by considering the mechanisms that might cause the change or effect and the implied form of the change that would be expected. In addition, the identification may be aided by direct inspection of the data to suggest the form of effect due to the known event, and supplementary evidence may sometimes be available from examination of the residuals from a model fitted before the intervention term is introduced.

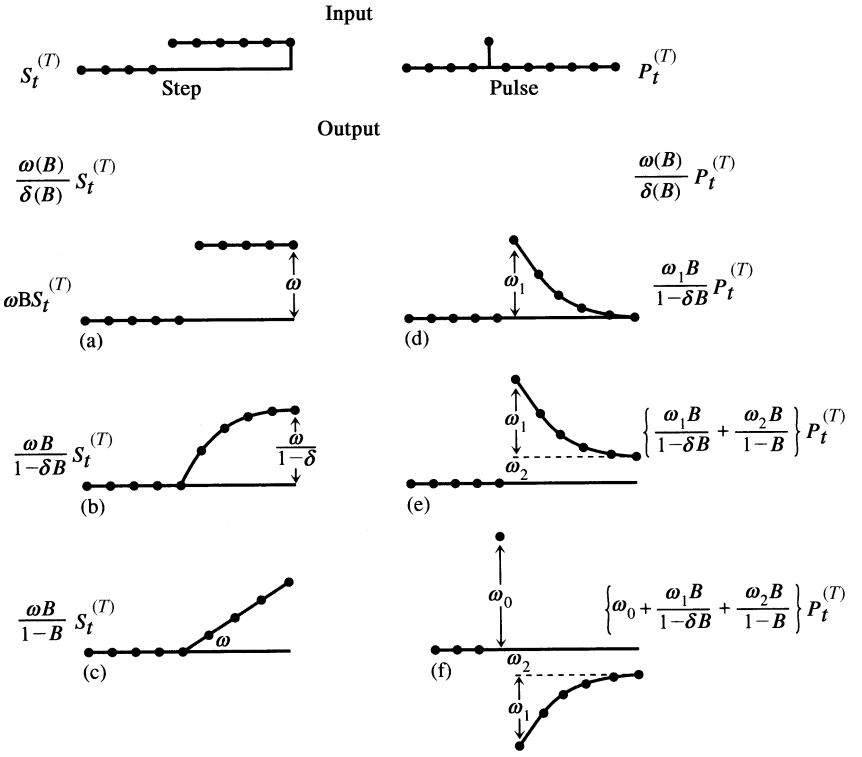

FIGURE 13.1 Responses to a step and a pulse input: (a–c) Response to a step input for various simple transfer function models, and (d–f) Response to a pulse input for some common models of interest.

Response Patterns Useful in Intervention Analysis. Several different response patterns

$$\mathcal{Y}_t = \delta^{-1}(B)\omega(B)B^b\xi_t$$

are possible through different choices of the transfer function. Figure 13.1 shows the responses for various simple transfer functions with both step and pulse indicators as input. For example, the model $\mathcal{Y}_t = \omega B S_t^{(T)}$ in Figure 13.1(a) can be used to represent a permanent step change in level of unknown magnitude ω after time T, while the form

$$\mathcal{Y}_t = \frac{\omega B}{1-\delta B} S_t^{(T)} \qquad 0 < \delta < 1 \tag{13.1.4}$$

in Figure 13.1(b), which implies that $\mathcal{Y}_t = \omega(1-\delta^{t-T})/(1-\delta)$, $t \geq T$, corresponds to a gradual change with rate δ that eventually approaches the long-run change in level equal to $\omega/(1-\delta)$. Similarly, the model

$$\mathcal{Y}_t = \frac{\omega_1 B}{1-\delta B} P_t^{(T)} \quad 0 < \delta < 1 \tag{13.1.5}$$

in Figure 13.1(d), which implies that $\mathcal{Y}_t = \omega_1 \delta^{t-T-1}, t > T$, would represent a sudden ''pulse'' change after time T of unknown magnitude ω_1, followed by a gradual decay of

rate δ back to the original preintervention level with no permanent effect. More complex response patterns can be obtained by various linear combinations of the simpler forms, such as in the case of Figure 13.1(f). It is also noted that since $(1 - B)S_t^{(T)} = P_t^{(T)}$, any of the transfer function models that involve $S_t^{(T)}$ could equally well be represented in terms $P_t^{(T)}$.

The following additional points concerning the intervention models are worthy of note. The function $\mathcal{Y}_t$ represents the additional effect of the intervention event over the noise or ''background'' series N_t. Hence, when possible, the model $N_t = [\theta(B)/\varphi(B)]a_t$ for the noise is identified based on the usual procedures applied to the time series observations available before the date of the intervention, that is, Y_t, $t < T$. Also, it is assumed in model (13.1.1) that only the level of the series is affected by the intervention and, in particular, that the form and the parameters of the time series model for N_t are the same before and after the intervention. One should also recognize that there can be considerable differences in the accuracy with which the intervention model parameters can be estimated depending on whether the noise N_t is stationary or nonstationary, as well as on whether permanent or transitory effects are postulated.

In general, the parameter estimates and their standard errors for the intervention model

$$Y_t = \frac{\omega(B)B^b}{\delta(B)}\xi_t + \frac{\theta(B)}{\varphi(B)}a_t \tag{13.1.6}$$

are obtained by the least-squares method of estimation for transfer function–noise models, as described in Section 12.3. Diagnostic checking based on the residuals $\hat{a}_t$ from the fitted model can also be performed using methods similar to those previously employed to assess the adequacy of a fitted model.

13.1.2 Example of Intervention Analysis

Box and Tiao (1975) considered the monthly time series consisting of the rate of change in the U.S. consumer price index (CPI) for the period July 1953 through December 1972. Beginning in September 1971, phase I economic control went into effect for 3 months, and after that phase II was in effect. The problem was to investigate the possible effect of the phase I and II controls on the rate of change in the CPI.

Inspection of the sample autocorrelation functions of the rate of change of the CPI and its first differences for the 218 monthly observations prior to phase I suggested a noise model of the form

$$(1 - B)N_t = (1 - \theta B)a_t \tag{13.1.7}$$

with maximum likelihood estimates $\hat{\theta} = 0.84$ and $\hat{\sigma}_a = 0.0019$. Examination of the residuals and their autocorrelations reveals no obvious inadequacies in this model.

Then, to address the question of the possible effects of phase I and II controls, it is assumed that phase I and II are expected to produce changes in the level of the rate of change of the CPI, and that the form of the noise model remains the same. Based on these assumptions, the appropriate model to assess the impact of the controls is

$$Y_t = \omega_1\xi_{1t} + \omega_2\xi_{2t} + \frac{1 - \theta B}{1 - B}a_t \tag{13.1.8}$$

where

$$\xi_{1t} = \begin{cases} 1 & t = \text{September, October, or November 1971} \\ 0 & \text{otherwise} \end{cases}$$

$$\xi_{2t} = \begin{cases} 1 & t \geq \text{December 1971} \\ 0 & \text{otherwise} \end{cases}$$

The nonlinear least-squares estimates of the parameters in model (13.1.8) were obtained, with standard errors in parentheses, as

$$\hat{\theta} = 0.85(0.05) \qquad \hat{\omega}_1 = -0.0022(0.0010) \qquad \hat{\omega}_2 = -0.0008(0.0009)$$

Hence, the analysis suggests that a drop in the rate of increase of the CPI is associated with phase I, but the effect of phase II is much less certain.

Many other examples of the use of intervention analysis have appeared in the literature. These include studies of the effects of regulations for engine design changes in new cars on oxidant pollution levels in the Los Angeles area (Box and Tiao, 1975), the effect of a change in policy in relation to debt collection on bad debt collections (Jenkins, 1979), the effectiveness of seat belt legislation on road deaths (Bhattacharyya and Layton, 1979), and the impact of the Arab oil embargo on electricity consumption in the United States (Montgomery and Weatherby, 1980).

13.1.3 Nature of the ML Estimator for a Simple Level Change Model

It is instructive to consider the nature of the maximum likelihood estimator of the intervention parameters, such as those in (13.1.8), for some relatively simple situations. We consider the simple model

$$Y_t = \omega \xi_t + N_t \tag{13.1.9}$$

where $N_t = \phi^{-1}(B)\theta(B)a_t$. This model can be written, formally, as

$$\pi(B)Y_t = \omega \pi(B)\xi_t + a_t \tag{13.1.10}$$

where $\pi(B) = \theta^{-1}(B)\phi(B) = 1 - \sum_{i=1}^{\infty} \pi_i B^i$. Letting $w_t = \pi(B)Y_t$ and $x_t = \pi(B)\xi_t$, we can write (13.1.10) in the form of a simple linear model $w_t = \omega x_t + a_t$, $t = 1, 2, \ldots, n$. Hence, the maximum likelihood estimator of ω is approximately

$$\hat{\omega} = \frac{\sum_{t=1}^{n} x_t w_t}{\sum_{t=1}^{n} x_t^2} \tag{13.1.11}$$

with $\text{var}[\hat{\omega}] = \sigma_a^2 / \sum_{t=1}^{n} x_t^2$.

Example with a Step Change Input and Nonstationary Noise. Let us consider a special case of (13.1.9) where $\xi_t = BS_t^{(T)}$ represents a step change after time T. Then, $x_t = \pi(B)BS_t^{(T)} = 1 - \sum_{i=1}^{t-T-1} \pi_i$, $t > T + 1$, with $x_{T+1} = 1$ and $x_t = 0$ for $t \leq T$. For the discussion that follows, we suppose that n is large, and that a relatively large number of observations are available before and after the intervention time T.

Now suppose that the noise N_t in (13.1.9) is nonstationary with generalized autoregressive operator $\varphi(B) = \phi(B)(1 - B)$ so that

$$\pi(B) = \theta^{-1}(B)\phi(B)(1 - B) = \tilde{\pi}(B)(1 - B)$$

with $\tilde{\pi}(B) = \theta^{-1}(B)\phi(B) = 1 - \sum_{j=1}^{\infty} \tilde{\pi}_j B^j$. Then, $x_t = \tilde{\pi}(B)B(1 - B)S_t^{(T)} = \tilde{\pi}(B)P_{t-1}^{(T)} = \tilde{\pi}_{t-T-1}$, $t \geq T + 1$, and hence

$$\sum_{t=1}^{n} x_t^2 = \sum_{t=T+1}^{n} \tilde{\pi}_{t-T-1}^2 \simeq \sum_{i=0}^{\infty} \tilde{\pi}_i^2 \equiv \eta_0$$

Also, $w_t = \pi(B)Y_t = Y_t - \overline{Y}_{t-1}$, where $\overline{Y}_{t-1} = \sum_{i=1}^{\infty} \pi_i Y_{t-i}$ is a weighted average of values prior to t (since $\sum_{i=1}^{\infty} \pi_i = 1$ when $d = 1$). Following the results in Box and Tiao (1975), it can then be shown that

$$\begin{aligned}\sum_{t=1}^{n} x_t w_t = \sum_{t=T+1}^{n} \tilde{\pi}_{t-T-1} w_t &\simeq \tilde{\pi}(B)\tilde{\pi}(F)(1 - B)Y_{T+1} \\ &= \sum_{s=0}^{\infty} \alpha_s Y_{T+1+s} - \sum_{s=0}^{\infty} \alpha_s Y_{T-s}\end{aligned}$$

where $a_s = \eta_s - \eta_{s+1}$ and the η_s are coefficients in $\tilde{\pi}(B)\tilde{\pi}(F) = \eta_0 + \sum_{s=1}^{\infty} \eta_s(B^s + F^s)$, such that $\sum_{s=0}^{\infty} \alpha_s = \eta_0 \equiv \sum_{i=0}^{\infty} \tilde{\pi}_i^2$. Therefore, in this situation, the maximum likelihood estimator of ω is

$$\hat{\omega} = \frac{\sum_{t=1}^{n} x_t w_t}{\sum_{t=1}^{n} x_t^2} \simeq (\eta_0)^{-1} \left(\sum_{s=0}^{\infty} \alpha_s Y_{T+1+s} - \sum_{s=0}^{\infty} \alpha_s Y_{T-s} \right) \tag{13.1.12}$$

with $\text{var}[\hat{\omega}] \simeq \sigma_a^2(\eta_0)^{-1}$. The estimator $\hat{\omega}$ can thus be interpreted as a contrast between two weighted moving averages, one consisting of the observations after the intervention and the other for the observations before the intervention, where the weights (α_s/η_0) are symmetrical.

For example, consider the case where N_t follows the IMA(0, 1, 1) model, $(1 - B)N_t = (1 - \theta B)a_t$, so that $\tilde{\pi}(B) = (1 - \theta B)^{-1}$ with $\tilde{\pi} = \theta^i$, $i \geq 1$. Then, $\eta_s = \theta^s/(1 - \theta^2)$, $s = 0, 1, \ldots$, and so $\alpha_s = (\theta^s - \theta^{s+1})/(1 - \theta^2) = \theta^s/(1 + \theta)$. Hence, the estimator in (13.1.12) becomes

$$\hat{\omega} \simeq (1 - \theta)^{-1} \left(\sum_{s=0}^{\infty} \theta^s Y_{T+1+s} - \sum_{s=0}^{\infty} \theta^s Y_{T-s} \right) \tag{13.1.12a}$$

with $\text{var}[\hat{\omega}] \simeq \sigma_a^2(1 - \theta^2)$. The estimator $\hat{\omega}$ is thus a contrast between two *exponentially* weighted moving averages, one consisting of the observations after the intervention and the other for the observations before the intervention.

Now, as a second case, suppose that the noise instead follows the ARIMA(1, 1, 0) model, so that $\tilde{\pi}(B) = (1 - \phi B)$ with $\tilde{\pi}_1 = -\phi$ and $\tilde{\pi}_i = 0$ for $i > 1$. Then $\eta_0 = 1 + \phi^2$, $\eta_1 = -\phi$, and $\eta_s = 0$, $s > 1$. Hence, $\sum_{t=1}^{n} x_t^2 = 1 + \phi^2 = \eta_0$, $\alpha_0 = 1 + \phi + \phi^2$, $\alpha_1 = -\phi$, and it

follows that

$$\sum_{t=1}^{n} x_t w_t = (1-\phi B)(1-\phi F)(1-B)Y_{T+1}$$
$$= [(1+\phi+\phi^2)Y_{T+1} - \phi Y_{T+2}] - [(1+\phi+\phi^2)Y_T - \phi Y_{T-1}]$$

Thus, for this case we have

$$\hat{\omega} = \frac{\sum_{t=1}^{n} x_t w_t}{\sum_{t=1}^{n} x_t^2}$$
$$= (1+\phi^2)^{-1}\{[(1+\phi+\phi^2)Y_{T+1} - \phi Y_{T+2}]$$
$$- [(1+\phi+\phi^2)Y_T - \phi Y_{T-1}]\} \tag{13.1.13}$$

with $\text{var}[\hat{w}] = \sigma_a^2/(1+\phi^2)$. Again, the estimator $\hat{w}$ can be viewed as a contrast between two weighted averages of the same form, one of the postintervention observations Y_{T+1} and Y_{T+2} and the other of the preintervention observations Y_T and Y_{T-1}, but the weighted averages are only finite in extent because the noise model contains only an AR factor $(1-\phi B)$ and no MA factor as in the previous case.

Comparison with a Case with Stationary Noise. Finally, we consider a simpler situation of model (13.1.9), in which the noise is *stationary*, for example, an AR(1) model $(1-\phi B)N_t = a_t$. In this situation we obtain $x_t = (1-\phi B)BS_t^{(T)} = 1-\phi$ for $t > T+1$ with $x_{T+1} = 1$ and $w_t = (1-\phi B)Y_t = Y_t - \phi Y_{t-1}$. Then, it readily follows that

$$\hat{\omega} = \frac{\sum_{t=1}^{n} x_t w_t}{\sum_{t=1}^{n} x_t^2}$$
$$\simeq \frac{(1-\phi)\sum_{t=T+1}^{n}(Y_t - \phi Y_{t-1})}{(n-T)(1-\phi)^2} \simeq \overline{Y}_2 \tag{13.1.14}$$

where $\overline{Y}_2 = (n-T)^{-1}\sum_{t=T+1}^{n} Y_t$ denotes an unweighted average of all observations after the intervention, with $\text{var}[\hat{\omega}] = \sigma_a^2/[1+(n-T-1)(1-\phi)^2] \simeq \sigma_a^2/[(n-T)(1-\phi)^2]$. Notice that because of the stationarity of the noise, we have an *unweighted* average of postintervention observations and also that there is no adjustment for the preintervention observations because they are assumed to be *stationary* about a *known mean of zero*. Also note that in the stationary case, the variance of $\hat{\omega}$ decreases proportionally with $1/(n-T)$, whereas in the previous nonstationary noise situations, $\text{var}[\hat{\omega}]$ is essentially a constant not dependent on the sample size. This reflects the differing degrees of accuracy in the estimators of intervention model parameters, such as the level shift parameter ω, that can be expected in large samples between the nonstationary noise and the stationary noise model situations.

Specifically, in the model (13.1.9), with $\xi_t = BS_t^{(T)}$ equal to a step input, suppose that the noise process N_t is nonstationary ARIMA with $d = 1$, so that $\phi(B)(1-B)N_t = \theta(B)a_t$. Then, by applying the differencing operator $(1-B)$, the model

$$Y_t = \omega BS_t^{(T)} + N_t \tag{13.1.15}$$

can also be expressed as

$$y_t = \omega B P_t^{(T)} + n_t \tag{13.1.16}$$

where $y_t = (1 - B)Y_t$ and $n_t = (1 - B)N_t$, and hence n_t is a stationary ARMA(p, q) process. Therefore, the MLE of ω for the original model (13.1.15) with a (permanent) *step* input effect and *nonstationary* noise ($d = 1$) will have features similar to the MLE in the model (13.1.16), which has a (transitory) *pulse* input effect and *stationary* noise.

Of course, the model (13.1.9) can be generalized to allow for an unknown nonzero mean ω_0 before the intervention, $Y_t = \omega_0 + \omega\xi_t + N_t$, with $\xi_t = BS_t^{(T)}$, so that ω represents the *change in mean level* after the intervention. Then, for the stationary AR(1) noise model case, for example, similar to (13.1.14), it can be shown that the MLE of ω is $\hat{\omega} \simeq \overline{Y}_2 - \overline{Y}_1$, where $\overline{Y}_1 = T^{-1}\sum_{t=1}^{T} Y_t$ denotes the sample mean of all preintervention observations.

13.2 OUTLIER ANALYSIS FOR TIME SERIES

Time series observations may sometimes be affected by isolated events, disturbances, or errors that create spurious effects in the series and result in unusual patterns in the observations that are not consistent with the overall behavior of the time series. Such unusual observations may be referred to as *outliers*. They may be the result of unusual external events such as strikes, sudden political or economic changes, unusual weather events, sudden changes in a physical system, and so on, or simply due to recording or gross errors in measurement. The presence of such outliers in a time series can have substantial effects on the behavior of sample autocorrelations, partial autocorrelations, estimates of ARMA model parameters, and forecasting, and can even affect the specification of the model. If the time of occurrence T of an event that results in the outlying behavior is known, the unusual effects can often be accounted for by the use of intervention analysis techniques discussed in Section 13.1. However, since in practice the presence of outliers is often not known at the start of the analysis, additional procedures for detection of outliers and assessment of their possible impacts are important. In this section we discuss some useful models for representing outliers and corresponding methods, similar to the methods of intervention analysis, for detection of outliers. Some relevant references that deal with the topics of outlier detection, influence of outliers, and robust methods of estimation include Bruce and Martin (1989), Chang et al. (1988), Chen and Liu (1993), Martin and Yohai (1986), and Tsay (1986).

13.2.1 Models for Additive and Innovational Outliers

Following the work of Fox (1972), we consider two simple intervention models to represent two different types of outliers that might occur in practice. These are the *additive outlier* (AO) and the *innovational outlier* (IO) models. Let z_t denote the underlying time series process that is free of the impact of outliers, and let Y_t denote the observed time series. We assume that z_t follows the ARIMA(p, d, q) model $\varphi(B)z_t = \theta(B)a_t$. Then, an additive outlier at time T, or ''observational outlier,'' is modeled as

$$Y_t = \omega P_t^{(T)} + z_t = \omega P_t^{(T)} + \frac{\theta(B)}{\varphi(B)}a_t \tag{13.2.1}$$

where $P_t^{(T)} = 1$ if $t = T$, $P_t^{(T)} = 0$ if $t \neq T$, denotes the pulse indicator at time T. An innovational outlier at time T, or ''innovational shock,'' is modeled as

$$Y_t = \frac{\theta(B)}{\varphi(B)}(\omega P_t^{(T)} + a_t) = \omega \frac{\theta(B)}{\varphi(B)} P_t^{(T)} + z_t \tag{13.2.2}$$

Hence, an AO affects the level of the observed time series only at time T, $Y_T = \omega + z_T$, by an unknown additive amount ω, while an IO represents an extraordinary random shock at time T, $a_T + \omega = a_T^*$, which affects all succeeding observations $Y_T, Y_{T+1}, \ldots$ through the dynamics of the system described by $\psi(B) = \theta(B)/\varphi(B)$, such that $Y_t = \omega\psi_i + z_t$ for $t = T + i \geq T$. For a stationary series, the effect of the IO is temporary since ψ_i decay exponentially to 0, but for nonstationary series with $d \geq 1$, there can be permanent effects that approach a level shift or even ramp effect since ψ_i do not decay to 0. More generally, an observed time series Y_t might be affected by outliers of different types at several points of time $T_1, T_2, \ldots, T_k$, and the multiple outlier model of the following general form

$$Y_t = \sum_{j=1}^{k} \omega_j \nu_j(B) P_t^{(T_j)} + z_t \tag{13.2.3}$$

could be considered for use, where $\nu_j(B) = 1$ for an AO at time T_j and $\nu_j(B) = \theta(B)/\varphi(B)$ for an IO at time T_j. Problems of interest associated with these outlier models are to identify the timing and the type of outliers and to estimate the magnitude ω of the outlier effect, so that the analysis of the time series will adjust for these outlier effects.

Tsay (1988), Chen and Tiao (1990), and Chen and Liu (1993), among others, also consider allowance in (13.2.3) for level shift type of outlier effect at unknown time of the form $\omega S_t^{(T)}$. The occurrence of such an effect is often encountered in series where the underlying process z_t that is nonstationary, and such that there is a factor $(1 - B)$ in the AR operator $\varphi(B)$ of the ARIMA model for z_t. Then recall that $(1 - B)S_t^{(T)} = P_t^{(T)}$ so that a level shift type of outlier effect for the nonstationary observed series Y_t is equivalent to an AO effect for the first differenced series $(1 - B)Y_t$.

13.2.2 Estimation of Outlier Effect for Known Timing of the Outlier

We first consider the estimation of the impact ω of an AO in (13.2.1) and that of an IO in (13.2.2), respectively, in the situation where the parameters of the time series model for the underlying process z_t are assumed known. To motivate iterative procedures that have been proposed for the general case, it will also be assumed that the timing T of the outlier is given.

Let $\pi(B) = \theta^{-1}(B)\varphi(B) = 1 - \sum_{i=1}^{\infty} \pi_i B^i$ and define $e_t = \pi(B)Y_t$ for $t = 1, 2, \ldots, n$, in terms of the observed series Y_t. Then we can write the above outlier models, (13.2.2) and (13.2.1), respectively, as

$$\text{IO}: \quad e_t = \omega P_t^{(T)} + a_t \tag{13.2.4a}$$

$$\text{AO}: \quad e_t = \omega\pi(B)P_t^{(T)} + a_t = \omega x_{1t} + a_t \tag{13.2.4b}$$

where for the AO model, $x_{1t} = \pi(B)P_t^{(T)} = -\pi_i$ if $t = T + i \geq T$, $x_{lt} = 0$ if $t < T$, with $\pi_0 = -1$. Thus, we see from (13.2.4) that the information about an IO is contained solely in the ''residual'' e_T at the particular time T, whereas that for an AO is spread over the

stretch of residuals $e_T, e_{T+1}, e_{T+2}, \ldots$, with generally decreasing weights $1, -\pi_1, -\pi_2, \ldots$, because the π_i are absolutely summable due to the invertibility of the MA operator $\theta(B)$. Equivalently, when an AO is present at time T, we can see that the residuals constructed from the observed series Y_t, for $t \geq T$, will be affected as $e_t = \pi(B)Y_t = a_t - \omega\pi_i$ for $t = T + i$. Hence, in the presence of an AO, a relatively high proportion of the constructed residuals could be influenced and distorted relative to the underlying white noise series a_t. Consequently, the presence of AOs that are unaccounted for typically tend to have a much more substantial adverse effect on estimates of the autocorrelations and parameters of the ARMA model for z_t compared to the presence of innovational outliers.

From least-squares principles, the least-squares estimator of the outlier impact ω in the IO model is simply the residual at time T,

$$\text{IO:} \quad \hat{\omega}_{I,T} = e_T \tag{13.2.5a}$$

with $\text{var}[\hat{w}_{I,T}] = \sigma_a^2$, while that in the AO model is the linear combination of $e_T, e_{T+1}, \ldots$,

$$\text{AO:} \quad \hat{\omega}_{A,T} = \frac{e_T - \sum_{i=1}^{n-T} \pi_i e_{T+i}}{\sum_{i=0}^{n-T} \pi_i^2} = \frac{\pi^*(F)e_T}{\tau^2} \tag{13.2.5b}$$

with $\text{var}[\hat{\omega}_{A,T}] = \sigma_a^2/\tau^2$, where $\tau^2 = \sum_{i=0}^{n-T} \pi_i^2$ and $\pi^*(F) = 1 - \pi_1 F - \pi_2 F^2 - \cdots - \pi_{n-T} F^{n-T}$. The notation in (13.2.5) reflects the fact that the estimates depend upon the time T. Note that in an underlying autoregressive model $\varphi(B)z_t = a_t$, since then $\pi^*(B) = \pi(B) = \varphi(B)$ for $T < n - p - d$, and $e_t = \varphi(B)Y_t$, in terms of the observations Y_t, the estimate $\hat{\omega}_{A,T}$ in (13.2.5b) can be written as

$$\hat{\omega}_{A,T} = \frac{\varphi(F)\varphi(B)Y_T}{\tau^2}$$

Since $\tau^2 \geq 1$, it is seen in general that $\text{var}[\hat{\omega}_{A,T}] \leq \text{var}[\hat{\omega}_{I,T}] = \sigma_a^2$, and in some cases $\text{var}[\hat{\omega}_{A,T}]$ can be much smaller than σ_a^2. For example, in an MA(1) model for z_t, the variance of $\hat{\omega}_{A,T}$ would be $\sigma_a^2(1 - \theta^2)/(1 - \theta^{2(n-T+1)}) \simeq \sigma_a^2(1 - \theta^2)$ when $n - T$ is large.

Significance tests for the presence of an outlier of type AO or IO at the given time T can be formulated as a test of $\omega = 0$ in either model (13.2.1) or (13.2.2), against $\omega \neq 0$. The likelihood ratio test criteria can be derived for both situations and essentially take the form of the standardized statistics

$$\lambda_{I,T} = \frac{\hat{\omega}_{I,T}}{\sigma_a} \quad \text{and} \quad \lambda_{A,T} = \frac{\tau\hat{\omega}_{A,T}}{\sigma_a} \tag{13.2.6}$$

respectively, for IO and AO types. Under the null hypothesis that $\omega = 0$, both statistics in (13.2.6) will have the standard normal distribution.

For the level-shift-type outlier model $Y_t = \omega S_t^{(T)} + z_t$, we have $e_t = \omega\pi(B)S_t^{(T)} + a_t$ and

$$\pi(B)S_t^{(T)} = \left[\frac{\pi(B)}{1 - B}\right] P_t^{(T)} \equiv \tilde{\pi}(B)P_t^{(T)}$$

with $\tilde{\pi}(B) = \pi(B)/(1-B) = 1 - \sum_{j=1}^{\infty} \tilde{\pi}_j B^j$. So it follows from the estimation results in (13.2.4b) and (13.2.5b) that the MLE of ω in the level shift model is $\hat{\omega}_{L,T} = \tilde{\pi}^*(F)e_T/\tilde{\tau}^2$ with

$$\tilde{\pi}^*(F) = 1 - \tilde{\pi}_1 F - \tilde{\pi}_2 F^2 - \cdots - \tilde{\pi}_{n-T} F^{n-T}$$

and $\tilde{\tau}^2 = 1 + \tilde{\pi}_1^2 + \cdots + \pi\tilde{\pi}_{n-T}^2$. When $d = 1$ in the ARIMA model, $\varphi(B) = \phi(B)(1-B)$ and $\tilde{\pi}(B) = \theta^{-1}(B)\phi(B)$, and, as discussed earlier, the results for this situation are the same as for the AO in terms of the model for the first differences:

$$(1-B)Y_t = \omega(1-B)S_t^{(T)} + (1-B)z_t = \omega P_t^{(T)} + \frac{\theta(B)}{\phi(B)}a_t$$

13.2.3 Iterative Procedure for Outlier Detection

In practice, the time T of a possible outlier as well as the model parameters are unknown. To address the problem of detection of outliers at unknown times, iterative procedures that are relatively convenient computationally have been proposed by Chang et al. (1988), Tsay (1986), and Chen and Liu (1993) to identify and adjust for the effects of outliers.

At the first stage of this procedure, the ARIMA model is estimated for the observed time series Y_t in the usual way, assuming that the series contains no outliers. The residuals $\hat{e}_t$ from the model are obtained as $\hat{e}_t = \theta^{-1}(B)\hat{\varphi}(B)Y_t = \hat{\pi}(B)Y_t$, and $\hat{\sigma}_a^2 = n^{-1}\sum_{t=1}^{n}\hat{e}_t^2$ is obtained. Then the statistics, as in (13.2.6),

$$\hat{\lambda}_{I,t} = \frac{\hat{\omega}_{I,t}}{\hat{\sigma}_a} \quad \text{and} \quad \hat{\lambda}_{A,t} = \frac{\hat{\tau}\hat{\omega}_{A,t}}{\hat{\sigma}_a}$$

are computed for each time $t = 1, 2, \ldots, n$, as well as

$$\hat{\lambda}_T = \max_t[\max(|\hat{\lambda}_{I,t}|, |\hat{\lambda}_{A,t}|)]$$

where T denotes the time when this maximum occurs. The possibility of an outlier of type IO is identified at time T if $\hat{\lambda} = |\hat{\lambda}_{I,T}| > c$, where c is a prespecified constant with typical values for c of 3.0, 3.5, or 4.0. The effect of this IO can be eliminated from the residuals by defining $\tilde{e}_T = \hat{e}_T - \hat{\omega}_{I,T} = 0$ at T. If $\hat{\lambda}_T = |\hat{\lambda}_{A,T}| > c$, the possibility of an AO is identified at T, and its impact is estimated by $\hat{\omega}_{A,T}$ as in (13.2.5b). The effect of this AO can be removed from the residuals by defining $\hat{e}_t = \hat{e}_t - \hat{\omega}_{A,T}\hat{\pi}(B)P_t^{(T)} = \hat{e}_t + \hat{\omega}_{A,T}\hat{\pi}_{t-T}$ for $t \geq T$. In either case, a new estimate $\tilde{\sigma}_a^2$ is computed from the modified residuals $\tilde{e}_t$.

If any outliers are identified, the modified residuals $\tilde{e}_t$ and modified estimate $\tilde{\sigma}_a^2$, but the same parameters $\hat{\pi}(B) = \hat{\theta}^{-1}(B)\hat{\varphi}(B)$, are used to compute new statistics $\hat{\lambda}_{I,t}$ and $\hat{\lambda}_{A,t}$. The preceding steps are then repeated until all outliers are identified. Suppose that this procedure identifies outliers at k time points $T_1, T_2, \ldots, T_k$. Then the overall outlier model, as in (13.2.3),

$$Y_t = \sum_{j=1}^{k} \omega_j \nu_j(B) P_t^{(T_j)} + \frac{\theta(B)}{\varphi(B)}a_t \tag{13.2.7}$$

is estimated for the observed series Y_t, where $v_j(B) = 1$ for an AO and $v_j(B) = \theta(B)/\varphi(B)$ for an IO at time T_j. A revised set of residuals

$$\hat{e}_t = \hat{\theta}^{-1}(B)\hat{\varphi}(B)\left[Y_t - \sum_{j=1}^{k} \hat{\omega}_j \hat{v}_j(B) P_t^{(T_j)}\right]$$

and a new $\hat{\sigma}_a^2$ are obtained from this fitted model. The previous steps of the procedure can then be repeated with new residuals, until all outliers are identified and a final model of the general form of (13.2.7) is estimated. If desired, a modified time series of observations in which the effects of the outliers have been removed can be constructed as $\tilde{z}_t = Y_t - \sum_{j=1}^{k} \hat{\omega}_j \hat{v}_j(B) P_t^{(T_j)}$.

The procedure above can be implemented, with few modifications, to any existing software capable of estimation of ARIMA and transfer function–noise models. An implementation in the R package will be demonstrated below. The technique can be a useful tool in the identification of potential time series outliers that if undetected could have a negative impact on the effectiveness of modeling and estimation. However, there should be some cautions concerning the systematic use of such ''outlier adjustment'' procedures, particularly with regard to the overall interpretation of results, the appropriateness of a general model specification for ''outliers'' such as (13.2.7), which treats the outliers as deterministic constants, and the possibilities for ''overspecification'' in the number of outliers. Whenever possible, it would always be highly desirable to search for the causes or sources of the outliers that may be identified by the foregoing procedure, so that the outlying behavior can be better understood and properly accounted for in the analysis. Also, although the foregoing procedures should perform well when the series has only a few relatively isolated outliers, there could be difficulties due to ''masking effects'' when the series has multiple outliers that occur in patches, especially when they are in the form of additive outliers and level shift effects. Modifications to the basic procedure to help remedy these difficulties associated with multiple outliers, including *joint* estimation of all identified outlier effects and the model parameters within the iteration stages, were proposed by Chen and Liu (1993).

13.2.4 Examples of Analysis of Outliers

We consider two numerical examples to illustrate the application of the outlier analysis procedures, discussed in the previous sections. For computational convenience, conditional least-squares estimation methods are used throughout in these examples.

Series D. The first example involves Series D, which represents ''uncontrolled" viscosity readings every hour from a chemical process. In Chapter 7, an AR(1) model $(1 - \phi B)z_t = \theta_0 + a_t$ has been suggested and fitted to this series. In the outlier detection procedure, the model is first estimated assuming that no outliers are present, and the results are given in Table 13.1(a). Then the AO and IO statistics as in (13.2.6) are computed for each time point t, using $\hat{\sigma}_a^2 = 0.08949$. Based on a critical value of $c = 3.5$, we lead to identification of an IO of rather large magnitude at time $T = 217$. The effect of this IO is removed by modifying the residual at T, a new estimate $\tilde{\sigma}_a^2 = 0.08414$ is obtained, and new outlier statistics are computed using $\tilde{\sigma}_a$. At this stage, no outliers are identified. Then, the time

TABLE 13.1 Outlier Detection and Parameter Estimation Results for Series C and D Examples

	Parameter[a]						Outlier			
	$\hat{\theta}_0$	$\hat{\phi}$	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{\omega}_3$	$\hat{\sigma}_a^2$	Time	$\hat{\omega}$	$\hat{\lambda}$	Type
(a) Series D										
Cycle 1	1.269	0.862				0.0895	217	−1.28	−4.29	IO
	(0.258)	(0.028)								
Final	1.181	0.872	−1.296			0.0841				
	(0.251)	(0.027)	(0.292)							
(b) Series C										
Cycle 1		0.813				0.0179	58	0.76	5.65	IO
		(0.038)					59	−0.51	−4.16	IO
							60	−0.44	−3.74	IO
Final		0.851	0.745	−0.551	−0.455	0.0132				
		(0.035)	(0.116)	(0.120)	(0.116)					

[a] Standard errors of parameter estimates are in parentheses.

series parameters and the outlier parameter ω in model (13.2.2), that is, in the model

$$Y_t = \frac{1}{1-\phi B}[\theta_0 + \omega P_t^{(T)} + a_t]$$

are estimated simultaneously, and the estimates are given in Table 13.1(a). Repeating the outlier detection procedure based on these new parameter estimates and corresponding residuals does not reveal any other outliers. Hence, only one extreme IO is identified, and adjusting for this IO does not result in much change in the estimate $\hat{\phi}$ of the time series model parameter, but gives about a 6% reduction in the estimate of σ_a^2. Several other potential outliers, at times $t = 29, 113, 115, 171, 268$, and 272, were also suggested during the outlier procedure as having values of the test statistics $\hat{\lambda}$ slightly greater than 3.0 in absolute value, but adjustment for such values did not affect the estimates of the model substantially.

Series C. The second example we consider is Series C, the ''uncontrolled'' temperature readings every minute in a chemical process. The model previously identified and fitted to this series is the ARIMA(1, 1, 0) model, $(1 - \phi B)(1 - B)z_t = a_t$. The estimation results for this model obtained assuming there are no outliers are given in Table 13.1(b). Proceeding with the sequence of calculations of the outlier test statistics and using the critical value of $c = 3.5$, we first identify an IO at time 58. The residual at time 58 is modified, we obtain a new estimate $\tilde{\sigma}_a^2 = 0.01521$, and next an IO at time 59 is identified. This residual is modified, a new estimate $\tilde{\sigma}_a^2 = 0.01409$ is obtained, and then another IO at time 60 is indicated. After this, no further outliers are identified. These innovational outliers at times 58, 59, and 60 are rather apparent in Figure 13.2(a), which shows a time series plot of the residuals from the initial model fit before any adjustment for outliers.

Then the time series outlier model

$$(1 - B)Y_t = \frac{1}{1-\phi B}[\omega_1 P_t^{(58)} + \omega_2 P_t^{(59)} + \omega_3 P_t^{(60)} + a_t]$$

is estimated for the series, and the results are presented in Table 13.1(b). The residuals are shown in Figure 13.2(b). No other outliers are detected when the outlier procedure is repeated with the new model parameter estimates. In this example we see that adjustment for the outliers has a little more effect on the estimate $\hat{\phi}$ of the time series parameter than in the previous case, and it reduces the estimate of σ_a^2 substantially by about 26%. Figure 13.2(b) clearly shows the reduction in variability due to the outlier adjustment.

Calculations Using R. The detection and adjustment for outliers in time series can be performed using the TSA package in R. The code needed to do the analysis for Series C and D is as follows:

```
> library(TSA)
> m1.C=arima(seriesC,order=c(1,1,0))
> m1.C
> detectAO(m1.C); detectIO(m1.C)
> m2.C=arimax(seriesC,order=c(1,1,0),io=c(58,59,60))
> m2.C
```

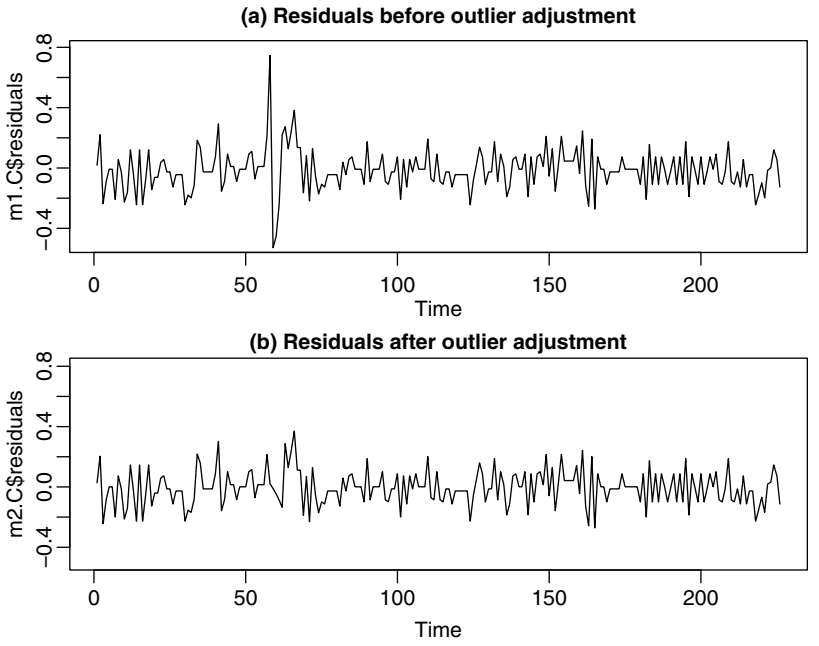

FIGURE 13.2 Residuals from the ARIMA(1, 1, 0) model fitted to Series C before and after adjustment for innovational outliers at $t = 58$, 59, and 60.

```
> m1.D=arima(seriesD,order=c(1,0,0))
> m1.D
> detectAO(m1.D); detectIO(m1.D)
> m2.D=arimax(seriesD,order=c(1,0,0),io=c(217))
> m2.D
```

Figure 13.2 that shows the residuals for Series C before and after the outlier adjustment can be reproduced in R as follows:

```
> par(mfrow=c(2,1))
> plot(m1.C$residuals,ylim=c(-0.5,0.8),
       main='(a) Residuals before outlier adjustment')
> plot(m2.C$residuals,ylim=c(-0.5,0.8),
       main='(b) Residuals after outlier adjustment')
```

13.3 ESTIMATION FOR ARMA MODELS WITH MISSING VALUES

In some situations in practice, the values of a time series z_t may not be observed at equally spaced times because there may be ''missing values'' corresponding to certain time points. In this section we discuss briefly the maximum likelihood estimation of parameters in an ARIMA(p, d, q) model for such situations, through consideration of the calculation of the exact Gaussian likelihood function for the observed data. It is shown that for series with missing observations, the likelihood function can conveniently be constructed using the

state-space form of the model and associated Kalman filtering procedures, as discussed in Sections 5.5 and 7.4, but modified to accommodate the missing data. These methods for evaluation of the likelihood in cases of irregularly spaced observations have been examined by Jones (1980), Harvey and Pierse (1984), Ansley and Kohn (1983, 1985), and Wincek and Reinsel (1986), among others. We also address briefly the related issue of estimation of the missing values in the time series.

13.3.1 State-Space Model and Kalman Filter with Missing Values

We suppose n observations are available at integer times $t_1 < t_2 < \cdots < t_n$, not equally spaced, from an ARIMA(p, d, q) process, which follows the model $\phi(B)(1 - B)^d z_t = \theta(B)a_t$. From Section 5.5.1, the process z_t has the state-space formulation given by

$$Y_t = \mathbf{\Phi} Y_{t-1} + \mathbf{\Psi} a_t \tag{13.3.1}$$

with $z_t = \mathbf{H}Y_t = [1, 0, \ldots, 0]Y_t$, where Y_t is the r-dimensional state vector and $r = \max(p + d, q + 1)$. Let $\Delta_i = t_i - t_{i-1}$ denote the time difference between successive observations $z_{t_{i-1}}$ and z_{t_i}, $i = 2, \ldots, n$. By successive substitutions, Δ_i times, on the right-hand side of (13.3.1), we obtain

$$Y_{t_i} = \mathbf{\Phi}^{\Delta_i} Y_{t_{i-1}} + \sum_{j=0}^{\Delta_i - 1} \mathbf{\Phi}^j \mathbf{\Psi} a_{t_i - j} \equiv \mathbf{\Phi}_i^* Y_{t_{i-1}} + a_{t_i}^* \tag{13.3.2}$$

where $\mathbf{\Phi}_i^* = \mathbf{\Phi}^{\Delta_i}$ and $a_{t_i}^* = \sum_{j=0}^{\Delta_i - 1} \mathbf{\Phi}^j \mathbf{\Psi} a_{t_i - j}$, with

$$\operatorname{cov}\left[a_{t_i}^*\right] = \mathbf{\Sigma}_i = \sigma_a^2 \sum_{j=0}^{\Delta_i - 1} \mathbf{\Phi}^j \mathbf{\Psi}\mathbf{\Psi}' \mathbf{\Phi}'^j$$

Thus, (13.3.2) together with the observation equation $z_{t_i} = \mathbf{H}Y_{t_i}$ constitutes a state-space model form for the *observed* time series data $z_{t_1}, z_{t_2}, \ldots, z_{t_n}$.

Therefore, the Kalman filter recursive equations as in (5.5.6) to (5.5.9) can be directly employed to obtain the state predictors $\hat{Y}_{t_i|t_{i-1}}$ and their error covariance matrices $\mathbf{V}_{t_i|t_{i-1}}$. So we can obtain the predictors

$$\hat{z}_{t_i|t_{i-1}} = E[z_{t_i} | z_{t_{i-1}}, \ldots, z_{t_1}] = \mathbf{H}\hat{Y}_{t_i|t_{i-1}} \tag{13.3.3}$$

for the observations z_{t_i} based on the previous *observed* data and their error variances

$$\sigma_a^2 v_i = \mathbf{H}\mathbf{V}_{t_i|t_{i-1}}\mathbf{H}' = E[(z_{t_i} - \hat{z}_{t_i|t_{i-1}})^2] \tag{13.3.4}$$

readily from the recursive Kalman filtering procedure. More specifically, the updating equations (5.5.6) and (5.5.7) in this missing data setting take the form

$$\hat{Y}_{t_i|t_i} = \hat{Y}_{t_i|t_{i-1}} + \mathbf{K}_i(z_{t_i} - \mathbf{H}\hat{Y}_{t_i|t_{i-1}}) \tag{13.3.5}$$

with

$$\mathbf{K}_i = \mathbf{V}_{t_i|t_{i-1}}\mathbf{H}'[\mathbf{H}\mathbf{V}_{t_i|t_{i-1}}\mathbf{H}']^{-1} \tag{13.3.6}$$

while the prediction equations (5.5.8) are given by

$$\hat{\mathbf{Y}}_{t_i|t_{i-1}} = \mathbf{\Phi}_i^* \hat{\mathbf{Y}}_{t_{i-1}|t_{i-1}} = \mathbf{\Phi}^{\Delta_i} \hat{\mathbf{Y}}_{t_{i-1}|t_{i-1}} \qquad \mathbf{V}_{t_i|t_{i-1}} = \mathbf{\Phi}_i^* \mathbf{V}_{t_{i-1}|t_{i-1}} \mathbf{\Phi}_i^{*\prime} + \mathbf{\Sigma}_i \tag{13.3.7}$$

with

$$\mathbf{V}_{t_t|t_i} = [\mathbf{I} - \mathbf{K}_i \mathbf{H}] \mathbf{V}_{t_i|t_{i-1}} \tag{13.3.8}$$

Notice that the calculation of the prediction equations (13.3.7) can be interpreted as computation of the successive one-step-ahead predictions:

$$\hat{\mathbf{Y}}_{t_{i-1}+j|t_{i-1}} = \mathbf{\Phi} \hat{\mathbf{Y}}_{t_{i-1}+j-1|t_{i-1}}$$
$$\mathbf{V}_{t_{i-1}+j|t_{i-1}} = \mathbf{\Phi} \mathbf{V}_{t_{i-1}+j-1|t_{i-1}} \mathbf{\Phi}' + \sigma_a^2 \mathbf{\Psi} \mathbf{\Psi}'$$

for $j = 1, \ldots, \Delta_i$, without any updating since there are no observations available between the time points t_{i-1} and t_i to provide any additional information for updating.

Exact Likelihood Function with Missing Values. The exact likelihood for the vector of observations $\mathbf{z}' = (z_{t_1}, z_{t_2}, \ldots, z_{t_n})$ is obtained directly from the quantities in (13.3.3) and (13.3.4) because the joint density of $\mathbf{z}$ can be expressed as the product of the conditional densities of the z_{t_i}, given $z_{t_{i-1}}, \ldots, z_{t_1}$, for $i = 2, \ldots, n$, which are Gaussian with conditional means and variances given by (13.3.3) and (13.3.4). Hence, the joint density of the observations $\mathbf{z}$ can be expressed as

$$p(\mathbf{z}|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = \prod_{i=1}^{n} (2\pi\sigma_a^2 v_i)^{-1/2} \exp\left[-\frac{1}{2\sigma_a^2} \sum_{i=1}^{n} \frac{(z_{t_i} - \hat{z}_{t_i|t_{i-1}})^2}{v_i} \right] \tag{13.3.9}$$

In (13.3.9), the quantities $\hat{z}_{t_i|t_{i-1}}$ and $\sigma_a^2 v_i$ are directly determined from the recursive filtering calculations (13.3.5)–(13.3.8). In the case of a stationary ARMA(p, q) model, the initial conditions required to start the filtering procedure can be determined readily (see, for example, Jones, (1980) and Section 5.5.2). However, for the nonstationary ARIMA model situation, some additional assumptions need to be specified concerning the process and the initial conditions. Appropriate methods for such cases have been examined by Ansley and Kohn (1985).

As a simple example to illustrate the missing data methods, consider the stationary AR(1) model $(1 - \phi B) z_t = a_t$. Then, (13.3.2) directly becomes (see, for example, Reinsel and Wincek, 1987)

$$z_{t_i} = \phi^{\Delta_i} z_{t_{i-1}} + \sum_{j=0}^{\Delta_{i-1}} \phi^j a_{t_i - j} \tag{13.3.10}$$

and it is readily determined that

$$\hat{z}_{t_i|t_{i-1}} = \phi^{\Delta_i} z_{t_{i-1}} \text{ and } \sigma_i^2 = \sigma_a^2 v_i = \frac{\sigma_a^2 (1 - \phi^{2\Delta_i})}{1 - \phi^2} \tag{13.3.11}$$

Hence, the likelihood for the observed data in the first-order autoregressive model with missing values is as given in (13.3.9), with these expressions for $\hat{z}_{t_i|t_{i-1}}$ and $\sigma_a^2 v_i$.

13.3.2 Estimation of Missing Values of an ARMA Process

A related problem of interest that often arises in the context of missing values for time series is that of estimating the missing values. Studies based on interpolation of missing values for ARIMA time series from a least-squares viewpoint were performed by Brubacher and Tunnicliffe Wilson (1976), Damsleth (1980), and Abraham (1981). Within the framework of the state-space formulation, estimates of missing values and their corresponding error variances can be derived conveniently through the use of recursive smoothing methods associated with the Kalman filter, which were discussed briefly in Section 5.5.3 and are described in general terms in Anderson and Moore (1979), for example. These methods have been considered more specifically for the ARIMA model with missing values by Harvey and Pierse (1984) and by Kohn and Ansley (1986).

For the special case of a pure autoregressive model, $\phi(B)z_t = a_t$, some rather simple and explicit interpolation results are available. For example, in an AR(p) process with a single missing value at time T surrounded by at least p consecutive observed values both before and after time T, it is well known (see, for example, Brubacher and Tunnicliffe Wilson, 1976) that the optimal interpolation of the missing value z_T is given by

$$\hat{z}_T = -d_0^{-1} \sum_{j=1}^{p} d_j (z_{T-j} + z_{T+j}) \tag{13.3.12}$$

where $d_j = \sum_{i=j}^{p} \phi_i \phi_{i-j}$, $\phi_0 = -1$, and $d_0 = 1 + \sum_{i=1}^{p} \phi_i^2$, with $E[(z_T - \hat{z}_T)^2] = \sigma_a^2 d_0^{-1} = \sigma_0^2 (1 + \sum_{i=1}^{p} \phi_i^2)^{-1}$. Notice that the value in (13.3.12) can be expressed as $\hat{z}_T = z_T - [\phi(F)\phi(B)z_T/d_0]$, with interpolation error equal to

$$\hat{e}_T = z_T - \hat{z}_T = \frac{\phi(F)\phi(B)z_T}{d_0} \tag{13.3.13}$$

As one way to establish the result (13.3.12), for convenience of discussion, suppose that z_T is the only missing value among times $t = 1, \ldots, n$, with $p + 1 \le T \le n - p$. Using a normal distribution assumption, the optimal (minimum MSE) estimate of z_T is $\hat{z}_T = E[z_T | z_1, \ldots, z_{T-1}, z_{T+1}, \ldots, z_n]$, which is also the best linear estimate without the normality assumption. Then, by writing the joint density of $\mathbf{z} = (z_1, \ldots, z_n)'$ in the form

$$p(z_1, \ldots, z_{T-1}, z_{T+1}, \ldots, z_n) p(z_T | z_1, \ldots, z_{T-1}, z_{T+1}, \ldots, z_n)$$

from basic properties of the multivariate normal distribution and its conditional distributions, it is easily deduced that the estimate $\hat{z}_T$, the conditional mean, is identical to the value of z_T that minimizes the ''sum-of-squares'' function in the exponent of the joint multivariate normal density of $\mathbf{z}$. Thus, since z_T occurs only in $p + 1$ terms of the exponent sum of squares, this reduces to finding the value of z_T to minimize $S = \sum_{i=0}^{p} a_{T+i}^2$, where

$a_t = z_t - \sum_{l=1}^{p} \phi_l z_{t-l}$. Now we obtain

$$\begin{aligned}\frac{\partial S}{\partial z_T} &= 2\left[\left(z_T - \sum_{l=1}^{p}\phi_l z_{T-l}\right) - \sum_{i=1}^{p}\phi_i\left(z_{T+i} - \sum_{l=1}^{p}\phi_l z_{T+i-l}\right)\right] \\ &= 2\left[\left(1 + \sum_{i=1}^{p}\phi_i^2\right)z_T + \sum_{i=0}^{p}\phi_i\left\{\sum_{l\neq i}^{p}\phi_l z_{T+i-l}\right\}\right] \\ &= 2\left[\left(1 + \sum_{i=1}^{p}\phi_i^2\right)z_T + \sum_{j=1}^{p}\left(\sum_{i=j}^{p}\phi_i\phi_{i-j}\right)(z_{T-j} + z_{T+j})\right]\end{aligned}$$

where $\phi_0 = -1$. Setting this partial derivative to zero and solving for z_T, we find that the estimate is given by $\hat{z}_T = -d_0^{-1}\sum_{j=1}^{p} d_j(z_{T-j} + z_{T+j})$, where $d_j = \sum_{i=j}^{p}\phi_i\phi_{i-j}$ and $d_0 = 1 + \sum_{i=1}^{p}\phi_i^2$. Notice that the estimate $\hat{z}_T$ can be seen to be determined from the solution for z_T to the relation $\phi(F)\phi(B)z_T = 0$, where $\phi(B) = 1 - \sum_{i=1}^{p}\phi_i B^i$ is the AR(p) operator. It can also be established that the error variance of the missing data estimate is given by $E[(z_T - \hat{z}_T)^2] = \sigma_a^2 d_0^{-1}$.

In the general ARMA model situation, Bruce and Martin (1989) and Ljung (1993), among others, have noted a close connection between the likelihood function construction in the case of missing values and the formulation of the consecutive data model likelihood with AOs specified for each time point that corresponds to a missing value. Hence, in effect, in such a time series AO model for consecutive data, for given values of the ARMA model parameters, the estimate of the outlier effect parameter ω corresponds to the interpolation error in the missing data situation. For example, in the autoregressive model situation, compare the result in (13.3.13) with the result given following (13.2.5b) for the AO model. Specifically, since $\pi(B) = \phi(B)$ in the AR(p) model, $e_T = \phi(B)Y_T$ and the estimate in (13.2.5b) reduces to $\hat{\omega}_{A,T} = [\phi(F)\phi(B)Y_T]/d_0 \equiv Y_T - \hat{Y}_T = \hat{e}_T$, the interpolation error given in (13.3.13). Furthermore, the sum-of-squares function in the likelihood (13.3.9) for the missing data situation is equal to the sum of squares obtained from a complete set of consecutive observations in which an AO has been assumed at each time point where a missing value occurs and for which the likelihood is evaluated at the maximum likelihood estimates for each of the corresponding AO effect parameters ω, for given values of the time series model parameters ϕ and θ. As an illustration, for the simple AR(1) model situation with a single isolated missing value at time T, from (13.3.11) the relevant term in the missing data sum-of-squares function is

$$\begin{aligned}\frac{(z_{T+1} - \phi^2 z_{T-1})^2}{1+\phi^2} &\equiv [(z_T - \hat{\omega}) - \phi z_{T-1}]^2 + [z_{T+1} - \phi(z_T - \hat{\omega})]^2 \\ &= (\hat{z}_T - \phi z_{T-1})^2 + (z_{T+1} - \phi\hat{z}_T)^2 \qquad (13.3.14)\end{aligned}$$

where

$$\hat{\omega} = z_T - \frac{\phi}{1+\phi^2}(z_{T-1} + z_{T+1}) = z_T - \hat{z}_T$$

is the maximum likelihood estimate of the outlier effect ω in the AO model (13.2.1), and the latter expressions in (13.3.14) represent the sum-of-squares terms in the consecutive data situation but with an AO modeled at time T.

Treating missing data as additive outliers does have an impact on estimation of the ARMA model parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, however, and ML estimates of these parameters in the missing data case are not identical to estimates that maximize a complete data likelihood for which an AO has been assumed at each time point where a missing value occurs. In fact, Basu and Reinsel (1996) established that MLEs of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ for the missing data situation are the same as estimates obtained from a model that assumes complete data with an AO at each time point where a missing value occurs when the method of *restricted maximum likelihood* estimation (e.g., as discussed in Section 9.5.2) is employed for this latter model formulation. We provide the following argument to establish this result.

Connection Between Exact Likelihood Function for Missing Data Situation and Restricted Likelihood. Let $\mathbf{z}_n = (z_{t_1}, z_{t_2}, \ldots, z_{t_n})'$ denote the $n \times 1$ vector of observations from the ARMA(p, q) process $\phi(B)z_t = \theta(B)a_t$ with $t_1 \equiv 1$ and $t_n = T$. Let $\mathbf{z}_0$ denote the $T \times 1$ vector consisting of the observations $\mathbf{z}_n$ with 0's inserted for values of times where observations are missing, and for convenience arrange as $\mathbf{z}_0 = (\mathbf{z}_n', 0')'$. Also, let $\mathbf{z} = (\mathbf{z}_n', \mathbf{z}_m')'$ denote the corresponding vector of (complete) values of the process, where $\mathbf{z}_m$ is the $m \times 1$ vector of the ''missing values,'' with $T = n + m$. We can write

$$\mathbf{z}_0 = \mathbf{X}\boldsymbol{\omega} + \mathbf{z} \tag{13.3.15}$$

where $\mathbf{X}$ is a $T \times m$ matrix with columns that are ''pulse'' unit vectors to indicate the m missing values, specifically, $\mathbf{X} = [\mathbf{0}, \mathbf{I}_m]'$ under the rearrangement of the data. Thus, (13.3.15) can be interpreted as a model that allows for AOs, with parameters $\boldsymbol{\omega}$, at all time points where a missing value occurs. Note that $\mathbf{z}_n = \mathbf{H}'\mathbf{z} \equiv \mathbf{H}'\mathbf{z}_0$ where $\mathbf{H}' = [\mathbf{I}_n, \mathbf{0}]$ is the $n \times T$ matrix whose rows are pulse unit vectors to indicate the n observed values.

From one perspective, (13.3.15) can be viewed as a ''regression model'' for the extended data vector $\mathbf{z}_0$ with $\boldsymbol{\omega}$ treated as unknown parameters and ARMA noise process $\{z_t\}$. (Note in fact that $\boldsymbol{\omega} = -\mathbf{z}_m$ by actual definition.) Let $\sigma_a^2\mathbf{V}_* = \text{cov}[\mathbf{z}]$ denote the $T \times T$ covariance matrix of the complete series of values. Then, the form of the restricted likelihood function for the extended data vector $\mathbf{z}_0$ under this regression model is given as in (9.5.11) of Section 9.5.2,

$$L_*\left(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2; \mathbf{z}_0\right) \propto \left(\sigma_a^2\right)^{-n/2} |\mathbf{V}_*|^{-1/2} |\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X}|^{-1/2} \times \exp\left[-\frac{1}{2\sigma_a^2}(\mathbf{z}_0 - \mathbf{X}\hat{\boldsymbol{\omega}})'\mathbf{V}_*^{-1}(\mathbf{z}_0 - \mathbf{X}\hat{\boldsymbol{\omega}})\right] \tag{13.3.16}$$

where $\hat{\boldsymbol{\omega}} = (\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{z}_0$. Recall from discussion in Section 9.5.2, however, that (13.3.16) has an equivalent representation as the density of the ''error contrast vector'' $\mathbf{H}'\mathbf{z}_0$, since $\mathbf{H}'$ is a full rank $(T - m) \times T$ matrix such that $\mathbf{H}'\mathbf{X} = \mathbf{0}$. Then noting that $\mathbf{H}'\mathbf{z}_0 = \mathbf{z}_n$, the observed data vector, expression (13.3.16) also represents the density of $\mathbf{z}_n$ and hence represents the exact likelihood based on the *observed* data vector $\mathbf{z}_n$, essentially by definition. However, we would now like to directly verify the equivalence between (13.3.16) and the exact likelihood (density) function of the observed data vector $\mathbf{z}_n$.

For this, we express the covariance matrix of $\mathbf{z} = (\mathbf{z}_n', \mathbf{z}_m')'$ in partitioned form as

$$\text{cov}[\mathbf{z}] = \sigma_a^2 \mathbf{V}_* = \sigma_a^2 \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix}$$

where $\sigma_a^2 \mathbf{V}_{11} = \text{cov}[\mathbf{z}_n]$ in particular. We let $\mathbf{V}^{ij}$, $i, j = 1, 2$, denote the block matrices of $\mathbf{V}_*^{-1}$ corresponding to the above partitioning of $\mathbf{V}_*$. Then using basic results for partitioned matrices (e.g., Rao (1965), or Appendix A7.1.1), we can readily derive that the restricted likelihood expression in (13.3.16) is the same as the likelihood (density) for the observed data vector $\mathbf{z}_n$. That is, from results on partitioned matrices, we first have that

$$\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X} \equiv \mathbf{V}^{22} = \left(\mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\right)^{-1} \tag{13.3.17}$$

and $|\mathbf{V}_*| = |\mathbf{V}_{11}||\mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}|$. Hence, the determinant factor in (13.3.16) is $|\mathbf{V}_*|^{-1/2}|\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X}|^{-1/2} = |\mathbf{V}_{11}|^{-1/2}$. Also, the quadratic form in (13.3.16) is expressible as

$$\begin{aligned} &\mathbf{z}_0'[\mathbf{V}_*^{-1} - \mathbf{V}_*^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_*^{-1}]\mathbf{z}_0 \\ &\quad = \mathbf{z}_n'[\mathbf{V}^{11} - \mathbf{V}^{12}(\mathbf{V}^{22})^{-1}\mathbf{V}^{21}]\mathbf{z}_n = \mathbf{z}_n'\mathbf{V}_{11}^{-1}\mathbf{z}_n \end{aligned}$$

again using a basic result on the inverse of a partitioned matrix. Therefore, expression (13.3.16) is equal to

$$p(\mathbf{z}_n) \propto \left(\sigma_a^2\right)^{-n/2} |\mathbf{V}_{11}|^{-1/2} \exp\left[-\frac{1}{2\sigma_a^2}\mathbf{z}_n'\mathbf{V}_{11}^{-1}\mathbf{z}_n\right] \tag{13.3.18}$$

which, since $\mathbf{z}_n$ is distributed as normal $N(\mathbf{0}, \sigma_a^2\mathbf{V}_{11})$, is the likelihood based on the observed data vector $\mathbf{z}_n$.

This equivalence establishes a device for obtaining ML estimates in ARMA models with missing values by using an REML estimation routine for the extended data vector $\mathbf{z}_0$ by setting up a regression component $\mathbf{X}\boldsymbol{\omega}$ that includes an indicator variable (AO term) for each missing observation. Estimation of the ''extended data'' regression model (13.3.15) with ARMA errors by the method of REML then results in ML estimates of the ARMA model parameters based on the observed data $\mathbf{z}_n$. Finally, we note that the GLS estimate of $\boldsymbol{\omega}$ in model (13.3.15) is

$$\begin{aligned} \hat{\boldsymbol{\omega}} &= (\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{z}_0 \\ &= (\mathbf{V}^{22})^{-1}\mathbf{V}^{21}\mathbf{z}_n \equiv -\mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{z}_n = -E[\mathbf{z}_m|\mathbf{z}_n] \end{aligned} \tag{13.3.19}$$

so the estimates of the missing values $\mathbf{z}_m$ are obtained as $\hat{\mathbf{z}}_m = -\hat{\boldsymbol{\omega}}$ immediately as a by-product of the fitting of the model (13.3.15), with estimation error covariance matrix $\text{cov}[\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}] \equiv \text{cov}[\mathbf{z}_m - \hat{z}_m] = \sigma_a^2(\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X})^{-1}$ directly available as well. In addition, for a complete data vector situation, if there were additive outliers specified at the given times corresponding to $\mathbf{z}_m$, then model (13.3.15) could be used to estimate ''smoothed values'' of the observations at all times where an AO is proposed to occur, $\hat{\mathbf{z}}_m = -\hat{\boldsymbol{\omega}}$ as

given in (13.3.19), and magnitudes of the outliers can be estimated by the differences between the observed values and the interpolated values, $\mathbf{z}_m - \hat{\mathbf{z}}_m$.

EXERCISES

13.1. In an analysis (Box and Tiao, 1975) of monthly data Y_t on smog-producing oxidant, allowance was made for two possible ''interventions'' I_1 and I_2 as follows:

I_1: In early 1960, diversion of traffic from the opening of the Golden State Freeway and the coming into effect of a law reducing reactive hydrocarbons in gasoline sold locally.

I_2: In 1966, the coming into effect of a law requiring all new cars to have modified engine design. In the case of this intervention, allowance was made for the well-known fact that the smog phenomenon is different in summer and winter months.

In a pilot analysis of the data, the following intervention model was used:

$$Y_t = \omega_1 \xi_{1t} + \frac{\omega_2}{1 - B^{12}} \xi_{2t} + \frac{\omega_3}{1 - B^{12}} \xi_{3t} + \frac{(1 - \theta B)(1 - \Theta B^{12})}{1 - B^{12}} a_t$$

where

$$\xi_{1t} = \begin{cases} 0 & t < \text{Jan. } 1960 \\ 1 & t \geq \text{Jan. } 1960 \end{cases} \qquad \xi_{2t} = \begin{cases} 0 & t < \text{Jan. } 1966 \\ 1 & t \geq \text{Jan. } 1966 \end{cases} \quad \text{(summer months)}$$

$$\xi_{3t} = \begin{cases} 0 & t < \text{Jan. } 1966 \\ 1 & t \geq \text{Jan. } 1966 \end{cases} \quad \text{(winter months)}$$

(a) Show that the model allows for the following:

(1) A possible step change in January 1960 of size ω_1, possibly produced by I_1.

(2) A ''staircase function'' of annual step size ω_2 to allow for possible summer effect of cumulative influx of cars with new engine design.

(3) A ''staircase function'' of annual step size ω_3 to allow for possible winter effect of cumulative influx of cars with new engine design.

(b) Describe what steps you would take to check the representational adequacy of the model.

(c) Assuming you were satisfied with the checking after (b), what conclusions would you draw from the following results? (Estimates are shown with their standard errors below in parentheses.)

$$\hat{\omega}_1 = -1.09 \quad \hat{\omega}_2 = -0.25 \quad \hat{\omega}_3 = -0.07 \quad \hat{\theta} = -0.24 \quad \hat{\Theta} = 0.55$$
$$(\pm 0.13) \qquad (\pm 0.07) \qquad (\pm 0.06) \qquad (\pm 0.03) \qquad (\pm 0.04)$$

(d) The data for this analysis are listed as Series R in the Collection of Time Series in Part Five. Use these data to perform your own intervention analysis.

13.2. A general transfer function model of the form

$$Y_t = \sum_{j=1}^{k} \delta_j^{-1}(B)\omega_j(B)\xi_{jt} + \phi^{-1}(B)\theta(B)a_t \equiv \mathcal{Y}_t + N_t$$

can include input variables ξ_j, which are themselves time series, and other inputs ξ_i, which are indicator variables. The latter can estimate (and eliminate) the effects of interventions of the kind described in Exercise 13.1 and, in particular, are often useful in the analysis of sales data.

Let $\xi_t^{(T)}$ be an indicator variable that takes the form of a unit pulse at time T, that is

$$\xi_t^{(T)} = \begin{cases} 0 & t \neq T \\ 1 & t = T \end{cases}$$

For illustration, consider the models

$$(1)\ \mathcal{Y}_t = \frac{\omega_1 B}{1 - \delta B}\xi_t^{(T)} \qquad (\text{with } \omega_1 = 1.0, \delta = 0.5)$$

$$(2)\ \mathcal{Y}_t = \left(\frac{\omega_1 B}{1 - \delta B} + \frac{\omega_2 B}{1 - B}\right)\xi_t^{(T)} \qquad (\text{with } \omega_1 = 1.0, \delta = 0.5, \omega_2 = 0.3)$$

$$(3)\ \mathcal{Y}_t = \left(\omega_0 + \frac{\omega_1 B}{1 - \delta B} + \frac{\omega_2 B}{1 - B}\right)\xi_t^{(T)} \qquad (\text{with } \omega_0 = 1.5, \omega_1 = -1.0, \delta = 0.5, \omega_2 = -0.5)$$

Compute recursively the response $\mathcal{Y}_t$ for each of these models at times $t = T, T + 1, T + 2, \ldots$ and comment on their possible usefulness in the estimation and/or elimination of effects due to such phenomena as advertising campaigns, promotions, and price changes.

13.3. Figure 13.2 shows the residuals before and after an outlier adjustment for the temperature data in Series C. Construct a similar graph for the viscosity data in Series D.

13.4. A time series defined as $z_t = 1000 \log_{10}(H_t)$, where H_t is the price of hogs recorded annually by the U.S. Census of Agriculture over the period 1867–1948, was considered in Exercise 6.6.

(a) Estimate the parameters of the model identified for this series. Perform diagnostic check to determine the adequacy of the fitted model.

(b) Are additive or innovational outliers present in this series?

(c) If outliers are found, perform the appropriate adjustments to the basic ARIMA model and evaluate the results.

13.5. Daily air quality measurements in New York, May–September 1973, are available in the data file ''airquality'' in the R datasets package. The file provides data on four air quality variables, including the solar radiation measured from 8 a.m. to 12 noon at Central Park. The solar radiation series has a few missing values.

(a) Assuming that an AR(1) is appropriate for the series, derive an expression for the conditional expectation of the missing values, given the available data.

(b) Repeat the derivation in part (a) assuming that an AR(2) model is appropriate for the series.

(c) How would you evaluate the AR assumptions and proceed to develop a suitable model for this series?

14

MULTIVARIATE TIME SERIES ANALYSIS

Multivariate time series analysis involves the use of stochastic models to describe and analyze the relationships among *several* time series. While the focus in most of the earlier chapters has been on univariate methods, we will now assume that k time series, denoted as $z_{1t}, z_{2t}, \ldots, z_{kt}$, are to be analyzed, and we let $\boldsymbol{Z}_t = (z_{1t}, \ldots, z_{kt})'$ denote the time series vector at time t, for $t = 0, \pm 1, \ldots$. Such multivariate processes are of interest in a variety of fields such as economics, business, the social sciences, earth sciences (e.g., meteorology and geophysics), environmental sciences, and engineering. For example, in an engineering setting, one may be interested in the study of the simultaneous behavior over time of current and voltage, or of pressure, temperature, and volume. In economics, we may be interested in the variations of interest rates, money supply, unemployment, and so on, while sales volume, prices, and advertising expenditures for a particular commodity may be of interest in a business context. Multiple time series of this type may be contemporaneously related, some series may lead other series, or there may exist feedback relationships between the series.

In the study of multivariate processes, a framework is needed for describing not only the properties of the individual series but also the possible cross relationships among the series. Two key purposes for analyzing and modeling the series jointly are:

1. To understand the dynamic relationships over time among the series.
2. To improve accuracy of forecasts for individual series by utilizing the additional information available from the related series in the forecasts for each series.

With these objectives in mind, we begin this chapter by introducing some basic concepts and tools that are needed for modeling multivariate time series. We then describe the vector autoregressive, or VAR, models that are widely used in applied work. The properties of

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

these models are examined and methods for model identification, parameter estimation, and model checking are described. This is followed by a discussion of vector moving average and mixed vector autoregressive–moving average models, along with associated modeling tools. A brief discussion of nonstationary unit-root models and cointegration among vector time series is also included. We find that most of the basic concepts and results from univariate time series analysis extend to the multivariate case. However, new problems and challenges arise in the modeling of multivariate time series due to the greater complexity of models and parametrizations in the vector case. Methods designed to overcome such challenges are discussed. For a more detailed coverage of various aspects of multivariate time series analysis, see for example, Reinsel (1997), Lütkepohl (2006), and Tsay (2014).

14.1 STATIONARY MULTIVARIATE TIME SERIES

Let $\boldsymbol{Z}_t = (z_{1t}, \ldots, z_{kt})', t = 0, \pm 1, \pm 2, \ldots$, denote a k-dimensional time series vector of random variables of interest. The choice of the univariate component time series z_{it} that are included in $\boldsymbol{Z}_t$ will depend on the subject matter area and an understanding of the system under study, but it is implicit that the component series will be interrelated both contemporaneously and across time lags. The representation and modeling of these dynamic interrelationships is of main interest in multivariate time series analysis. Similar to the univariate case, an important concept in the model representation and analysis, which enables useful modeling results to be obtained from a finite sample realization of the series, is that of stationarity.

The vector process $\{\boldsymbol{Z}_t\}$ is (strictly) *stationary* if the probability distributions of the random vectors $(\boldsymbol{Z}_{t_1}, \boldsymbol{Z}_{t_2}, \ldots, \boldsymbol{Z}_{t_m})$ and $(\boldsymbol{Z}_{t_1+l}, \boldsymbol{Z}_{t_2+l}, \ldots, \boldsymbol{Z}_{t_m+l})$ are the same for arbitrary times $t_1, t_2, \ldots, t_m$, all m, and all lags or leads $l = 0, \pm 1, \pm 2, \ldots$. Thus, the probability distribution of observations from a stationary vector process is invariant with respect to shifts in time. Hence, assuming finite first and second moments exist, for a stationary process we must have $E[\boldsymbol{Z}_t] = \boldsymbol{\mu}$, constant for all t, where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)'$ is the mean vector of the process. Also, the vectors $\boldsymbol{Z}_t$ must have a constant covariance matrix for all t, which we denote by $\boldsymbol{\Sigma}_z \equiv \boldsymbol{\Gamma}(0) = E[(\boldsymbol{Z}_t - \boldsymbol{\mu})(\boldsymbol{Z}_t - \boldsymbol{\mu})']$. A less stringent definition of second-order, or covariance stationarity will be provided below.

14.1.1 Cross-Covariance and Cross-Correlation Matrices

For a stationary process $\{\boldsymbol{Z}_t\}$ the covariance between z_{it} and $z_{j,t+l}$ must depend only on the lag l, not on time t, for $i, j = 1, \ldots, k, l = 0, \pm 1, \pm 2, \ldots$. Hence, similar to definitions used in Section 12.1.1, we define the cross-covariance between the series z_{it} and z_{jt} at lag l as

$$\gamma_{ij}(l) = \text{cov}[z_{it}, z_{j,t+l}] = E[(z_{it} - \mu_i)(z_{j,t+l} - \mu_j)]$$

and denote the $k \times k$ matrix of *cross-covariances* at lag l as

$$\boldsymbol{\Gamma}(l) = E[(\boldsymbol{Z}_t - \boldsymbol{\mu})(\boldsymbol{Z}_{t+l} - \boldsymbol{\mu})'] = \begin{bmatrix} \gamma_{11}(l) & \gamma_{12}(l) & \cdots & \gamma_{1k}(l) \\ \gamma_{21}(l) & \gamma_{22}(l) & \cdots & \gamma_{2k}(l) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{k1}(l) & \gamma_{k2}(l) & \cdots & \gamma_{kk}(l) \end{bmatrix} \tag{14.1.1}$$

for $l = 0, \pm 1, \pm 2, \ldots$. The corresponding *cross-correlations* at lag l are

$$\rho_{ij}(l) = \text{corr}[z_{it}, z_{j,t+l}] = \frac{\gamma_{ij}(l)}{\{\gamma_{ii}(0)\gamma_{jj}(0)\}^{1/2}}$$

with $\gamma_{ii}(0) = \text{var}[z_{it}]$. Thus, for $i = j$, $\rho_{ii}(l) = \rho_{ii}(-l)$ denotes the autocorrelation function of the ith series z_{it}, and for $i \neq j$, $\rho_{ij}(l) = \rho_{ji}(-l)$ denotes the cross-correlation function between the series z_{it} and z_{jt}. The $k \times k$ *cross-correlation* matrix $\boldsymbol{\rho}(l)$ at lag l, with (i, j)th element equal to $\rho_{ij}(l)$, is given by

$$\boldsymbol{\rho}(l) = \mathbf{V}^{-1/2}\boldsymbol{\Gamma}(l)\mathbf{V}^{-1/2} = \{\rho_{ij}(l)\} \tag{14.1.2}$$

for $l = 0, \pm 1, \pm 2, \ldots$, where $\mathbf{V}^{-1/2} = \text{diag}\{\gamma_{11}(0)^{-1/2}, \ldots, \gamma_{kk}(0)^{-1/2}\}$. Note that $\boldsymbol{\Gamma}(l)' = \boldsymbol{\Gamma}(-l)$ and $\boldsymbol{\rho}(l)' = \boldsymbol{\rho}(-l)$, since $\gamma_{ij}(l) = \gamma_{ji}(-l)$. In addition, the cross-covariance matrices $\boldsymbol{\Gamma}(l)$ and cross-correlation matrices $\boldsymbol{\rho}(l)$ are nonnegative definite, since

$$\text{var}\left[\sum_{i=1}^{n} \boldsymbol{b}_i' \boldsymbol{Z}_{t-i}\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{b}_i' \boldsymbol{\Gamma}(i-j)\boldsymbol{b}_j \geq 0$$

for all positive integers n and all k-dimensional constant vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$.

14.1.2 Covariance Stationarity

The definition of stationarity given above is usually referred to as strict or strong stationarity. In general, a process $\{\boldsymbol{Z}_t\}$ that possesses finite first and second moments and that satisfies the conditions that $E[\boldsymbol{Z}_t] = \boldsymbol{\mu}$ does not depend on t and $E[(\boldsymbol{Z}_t - \boldsymbol{\mu})(\boldsymbol{Z}_{t+l} - \boldsymbol{\mu})']$ depends only on l is referred to as *weak, second-order*, or *covariance stationary*. In this chapter, the term stationary will generally be used in this latter sense of weak stationarity. For a stationary vector process, the cross-covariance and cross-correlation matrices provide useful summary information on the dynamic interrelations among the components of the process. However, because of the higher dimensionality $k > 1$ of the vector process, the cross-correlation matrices generally have more complicated structures and can be much more difficult to interpret than the autocorrelation functions in the univariate case. In Sections 4.2-4.4, we will examine the covariance properties implied by vector autoregressive, moving average, and mixed autoregressive-moving average models.

14.1.3 Vector White Noise Process

The simplest example of a stationary vector process is the *vector white noise process*, which plays a fundamental role as a building block for general vector processes. The vector white noise process is defined as a sequence of random vectors $\ldots, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_t, \ldots$ with $\boldsymbol{a}_t = (a_{1t}, \ldots, a_{kt})'$, such that $E[\boldsymbol{a}_t] = \mathbf{0}$, $E[\boldsymbol{a}_t \boldsymbol{a}_t'] = \boldsymbol{\Sigma}$, and $E[\boldsymbol{a}_t \boldsymbol{a}_{t+l}'] = \mathbf{0}$, for $l \neq 0$. Hence, its covariance matrices $\boldsymbol{\Gamma}(l)$ are given by

$$\boldsymbol{\Gamma}(l) = E[\boldsymbol{a}_t \boldsymbol{a}_{t+l}'] = \begin{cases} \boldsymbol{\Sigma} & \text{for} \quad l = 0 \\ \mathbf{0} & \text{for} \quad l \neq 0 \end{cases} \tag{14.1.3}$$

The $k \times k$ covariance matrix $\boldsymbol{\Sigma}$ is assumed to be positive definite, since the dimension k of the process could be reduced otherwise. Sometimes, additional properties will be assumed for the $\boldsymbol{a}_t$, such as normality or mutual independence over different time periods.

14.1.4 Moving Average Representation of a Stationary Vector Process

A multivariate generalization of Wold's theorem states that if $\{\boldsymbol{Z}_t\}$ is a purely nondeterministic (i.e., $\boldsymbol{Z}_t$ does not contain a purely deterministic component process whose future values can be perfectly predicted from the past values) stationary process with mean vector $\boldsymbol{\mu}$, then $\boldsymbol{Z}_t$ can be represented as an *infinite vector moving average (MA) process*,

$$\boldsymbol{Z}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t-j} = \boldsymbol{\mu} + \boldsymbol{\Psi}(B)\boldsymbol{a}_t \qquad \boldsymbol{\Psi}_0 = \mathbf{I} \tag{14.1.4}$$

where $\boldsymbol{\Psi}(B) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j B^j$ is a $k \times k$ matrix in the backshift operator B such that $B^j \boldsymbol{a}_t = \boldsymbol{a}_{t-j}$ and the $k \times k$ coefficient matrices $\boldsymbol{\Psi}_j$ satisfy the condition $\sum_{j=0}^{\infty} \|\boldsymbol{\Psi}_j\|^2 < \infty$, where $\|\boldsymbol{\Psi}_j\|$ denotes the norm of $\boldsymbol{\Psi}_j$. The $\boldsymbol{a}_t$ form a vector white noise process with mean $\mathbf{0}$ and covariances given by (14.1.3). The covariance matrix of $\boldsymbol{Z}_t$ is then given by

$$\text{Cov}(\boldsymbol{Z}_t) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\Sigma} \boldsymbol{\Psi}_j'$$

The Wold representation in (14.1.4) is obtained by defining $\boldsymbol{a}_t$ as the error $\boldsymbol{a}_t = \boldsymbol{Z}_t - \hat{\boldsymbol{Z}}_{t-1}(1)$ of the best (i.e., minimum mean square error) one-step-ahead linear predictor $\hat{\boldsymbol{Z}}_{t-1}(1)$ of $\boldsymbol{Z}_t$ based on the infinite past $\boldsymbol{Z}_{t-1}, \boldsymbol{Z}_{t-2}, \ldots$. Thus, the $\boldsymbol{a}_t$ are mutually uncorrelated by construction since $\boldsymbol{a}_t$ is uncorrelated with $\boldsymbol{Z}_{t-j}$ for all $j \geq 1$ and, hence, is uncorrelated with $\boldsymbol{a}_{t-j}$ for all $j \geq 1$, and the $\boldsymbol{a}_t$ have a constant covariance matrix by stationarity of the process $\{\boldsymbol{Z}_t\}$. The best one-step-ahead linear predictor can be expressed as

$$\hat{\boldsymbol{Z}}_{t-1}(1) = \boldsymbol{\mu} + \sum_{j=1}^{\infty} \boldsymbol{\Psi}_j \{\boldsymbol{Z}_{t-j} - \hat{\boldsymbol{Z}}_{t-j-1}(1)\} = \boldsymbol{\mu} + \sum_{j=1}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t-j}$$

Consequently, the coefficient matrices $\boldsymbol{\Psi}_j$ in (14.1.4) have the interpretation of the linear regression matrices of $\boldsymbol{Z}_t$ on the $\boldsymbol{a}_{t-j}$ in that $\boldsymbol{\Psi}_j = \text{cov}[\boldsymbol{Z}_t, \boldsymbol{a}_{t-j}]\boldsymbol{\Sigma}^{-1}$.

In what follows, we will assume that $\boldsymbol{\Psi}(B)$ can be represented (at least approximately, in practice) as the product $\boldsymbol{\Phi}^{-1}(B)\boldsymbol{\Theta}(B)$, where $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Theta}(B)$ are finite autoregressive and moving average matrix polynomials of orders p and q, respectively. This leads to a class of linear models for vector time series $\boldsymbol{Z}_t$ defined by a relation of the form $\boldsymbol{\Phi}(B)(\boldsymbol{Z}_t - \mu) = \boldsymbol{\Theta}(B)\mathbf{a}_t$, or

$$(\boldsymbol{Z}_t - \boldsymbol{\mu}) - \sum_{j=1}^{p} \boldsymbol{\Phi}_j (\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}) = \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j} \tag{14.1.5}$$

A process $\{\boldsymbol{Z}_t\}$ is referred to as a *vector autoregressive--moving average*, or VARMA(p, q), process if it satisfies the relations (14.1.5) for a given white noise sequence $\{\boldsymbol{a}_t\}$.

We begin the discussion of this class of vector models by examining the special case when q is zero so that the process follows a pure vector autoregressive model of order p.

The discussion will focus on time-domain methods for analyzing vector time series and spectral methods will not be used. However, a brief summary of the spectral characteristics of stationary vector processes is provided in Appendix A14.1.

14.2 VECTOR AUTOREGRESSIVE MODELS

Among multivariate time series models, vector autoregressive models are the most widely used in practice. A major reason for this is their similarity to ordinary regression models and the relative ease of fitting these models to actual time series. For example, the parameters can be estimated using least-squares methods that yield closed-form expressions for the estimates. Other methods from multivariate regression analysis can be used at other steps of the analysis. Vector autoregressive models are widely used in econometrics, for example, to describe the dynamic behavior of economic and financial time series and to produce forecasts. This section examines the properties of vector autoregressive models and describes methods for order specification, parameter estimation, and model checking that can be used to develop these models in practice.

14.2.1 VAR(p) Model

A *vector autoregressive* model of order p, or VAR(p) model, is defined as

$$\boldsymbol{\Phi}(B)(\boldsymbol{Z}_t - \boldsymbol{\mu}) = \boldsymbol{a}_t$$

where $\boldsymbol{\Phi}(B) = \mathbf{I} - \boldsymbol{\Phi}_1 B - \boldsymbol{\Phi}_2 B^2 - \cdots - \boldsymbol{\Phi}_p B^p$, $\boldsymbol{\Phi}_i$ is a $k \times k$ parameter matrix, and $\boldsymbol{a}_t$ is a white noise sequence with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The model can equivalently be written as

$$(\boldsymbol{Z}_t - \boldsymbol{\mu}) = \sum_{j=1}^{p} \boldsymbol{\Phi}_j(\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}) + \boldsymbol{a}_t \tag{14.2.1}$$

The behavior of the process is determined by the roots of the determinantal equation $\det\{\boldsymbol{\Phi}(B)\} = 0$. In particular, the process is stationary if all the roots of this equation are greater than one in absolute value; that is, lie outside the unit circle (e.g., Reinsel,1997, Chapter 2). When this condition is met, $\{\boldsymbol{Z}_t\}$ has the infinite moving average representation

$$\boldsymbol{Z}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t-j} \tag{14.2.2}$$

or $\boldsymbol{Z}_t = \boldsymbol{\mu} + \boldsymbol{\Psi}(B)\boldsymbol{a}_t$, where $\boldsymbol{\Psi}(B) = \boldsymbol{\Phi}^{-1}(B)$ and the coefficient matrices $\boldsymbol{\Psi}_j$ satisfy the condition $\sum_{j=0}^{\infty} \|\boldsymbol{\Psi}_j\| < \infty$. Then, since $\boldsymbol{\Phi}(B)\boldsymbol{\Psi}(B) = \mathbf{I}$, the coefficient matrices can be calculated recursively from

$$\boldsymbol{\Psi}_j = \boldsymbol{\Phi}_1 \boldsymbol{\Psi}_{j-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{\Psi}_{j-p} \tag{14.2.3}$$

with $\boldsymbol{\Psi}_0 = \mathbf{I}$ and $\boldsymbol{\Psi}_j = \mathbf{0}$, for $j < 0$.

The moving average representation (14.2.2) is useful for examining the covariance properties of the process and it has a number of other applications. As in the univariate case, it is useful for studying forecast errors when the VAR(p) model is used for forecasting.

It is also used in *impulse response analysis* to determine how current or future values of the series are impacted by past changes or ''shocks'' to the system. The coefficient matrix $\boldsymbol{\Psi}_j$ shows the expected impact of a past shock $\boldsymbol{a}_{t-j}$ on the current value $\boldsymbol{Z}_t$. The response of a specific variable to a shock in another variable is often of interest in applied work. However, since the components of $\boldsymbol{a}_{t-j}$ are typically correlated, the individual elements of the $\boldsymbol{\Psi}_j$ can be difficult to interpret. To aid the interpretation, the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{a}_t$ can be diagonalized using a Cholesky decomposition $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}'$, where $\boldsymbol{L}$ is a lower triangular matrix with positive diagonal elements. Then, letting $\boldsymbol{b}_t = \boldsymbol{L}^{-1}\boldsymbol{a}_t$, we have $\mathrm{Cov}(\boldsymbol{b}_t) = \mathbf{I}_k$, and the model can be rewritten as

$$\boldsymbol{Z}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j^* \boldsymbol{b}_{t-j}$$

where $\boldsymbol{\Psi}_0^* = \boldsymbol{L}$ and $\boldsymbol{\Psi}_j^* = \boldsymbol{\Psi}_j \boldsymbol{L}$ for $j > 0$. The matrices $\boldsymbol{\Psi}_j^*$ are called the *impulse response weights* with respect to the orthogonal innovations $\boldsymbol{b}_t$. Since $\boldsymbol{L}$ is a lower triangular matrix, the ordering of the variables will, however, matter in this case. For further discussion and for applications of impulse response analysis, see Lütkepohl (2006, Chapter 2) and Tsay (2014, Chapter 2).

Reduced and Structural Forms. It is sometimes useful to express the VAR(p) process in (14.2.1) in the following slightly different form. Since the matrix $\boldsymbol{\Sigma} = E[\boldsymbol{a}_t \boldsymbol{a}_t']$ is assumed to be positive definite, there exists a lower triangular matrix $\boldsymbol{\Phi}_0^{\#}$ with ones on the diagonal such that $\boldsymbol{\Phi}_0^{\#}\boldsymbol{\Sigma}\boldsymbol{\Phi}_0^{\#\prime} = \boldsymbol{\Sigma}^{\#}$ is a diagonal matrix with positive diagonal elements. Hence, by premultiplying (14.2.1) by $\boldsymbol{\Phi}_0^{\#}$, we obtain the following representation:

$$\boldsymbol{\Phi}_0^{\#}(\boldsymbol{Z}_t - \boldsymbol{\mu}) = \sum_{j=1}^{p} \boldsymbol{\Phi}_j^{\#}(\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}) + \boldsymbol{b}_t \tag{14.2.4}$$

where $\boldsymbol{\Phi}_j^{\#} = \boldsymbol{\Phi}_0^{\#}\boldsymbol{\Phi}_j$ and $\boldsymbol{b}_t = \boldsymbol{\Phi}_0^{\#}\boldsymbol{a}_t$ with $\mathrm{Cov}[\boldsymbol{b}_t] = \boldsymbol{\Sigma}^{\#}$. This model displays the concurrent dependence among the components of $\boldsymbol{Z}_t$ through the lower triangular matrix $\boldsymbol{\Phi}_0^{\#}$ and is sometimes referred to as the *structural* form of the VAR(p) model. The model (14.2.1) that includes the concurrent relationships in the covariance matrix $\boldsymbol{\Sigma}$ of the errors and does not show them explicitly is referred to as the standard or *reduced form* of the VAR(p) model. Note that a diagonalizing transformation of this type was already used in the impulse response analysis described above, where the innovations $\boldsymbol{b}_t$'s were further normalized to have unit variance.

14.2.2 Moment Equations and Yule–Walker Estimates

For the VAR(p) model, the covariance matrices $\boldsymbol{\Gamma}(l) = \mathrm{Cov}(\boldsymbol{Z}_t, \boldsymbol{Z}_{t+l}) = \mathrm{Cov}(\boldsymbol{Z}_{t-l}, \boldsymbol{Z}_t) = E[(\boldsymbol{Z}_{t-l} - \boldsymbol{\mu})(\boldsymbol{Z}_t - \boldsymbol{\mu})']$ satisfy the matrix equations

$$\boldsymbol{\Gamma}(l) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(l-j)\boldsymbol{\Phi}_j' \tag{14.2.5}$$

for $l = 1, 2, \ldots$, with $\boldsymbol{\Gamma}(0) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(-j)\boldsymbol{\Phi}_j' + \boldsymbol{\Sigma}$. This result is readily derived using (14.2.1), noting that $E[(\boldsymbol{Z}_{t-l} - \boldsymbol{\mu})\boldsymbol{a}_{t-j}'] = \mathbf{0}$, for $j < l$. The matrix equations (14.2.5) are commonly referred to as the multivariate *Yule–Walker equations* for the VAR(p) model. For $l = 0, \ldots, p$, these equations can be used to solve for the $\boldsymbol{\Gamma}(l)$ simultaneously in terms of the AR parameter matrices $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Sigma}$.

Conversely, the AR coefficient matrices $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p$ and $\boldsymbol{\Sigma}$ can also be determined from the $\boldsymbol{\Gamma}$'s by first solving the Yule–Walker equations, for $l = 1, \ldots, p$, to obtain the parameters $\boldsymbol{\Phi}_j$. These equations can be written in matrix form as $\boldsymbol{\Gamma}_p \boldsymbol{\Phi}_{(p)} = \boldsymbol{\Gamma}_{(p)}$, with solution $\boldsymbol{\Phi}_{(p)} = \boldsymbol{\Gamma}_p^{-1} \boldsymbol{\Gamma}_{(p)}$, where

$$\boldsymbol{\Phi}_{(p)} = [\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p]' \qquad \boldsymbol{\Gamma}_{(p)} = [\boldsymbol{\Gamma}(1)', \ldots, \boldsymbol{\Gamma}(p)']'$$

and $\boldsymbol{\Gamma}_p$ is a $kp \times kp$ matrix with (i, j)th block of elements equal to $\boldsymbol{\Gamma}(i - j)$. Once the $\boldsymbol{\Phi}_j$ are determined from this, $\boldsymbol{\Sigma}$ can be obtained as

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(0) - \sum_{j=1}^{p} \boldsymbol{\Gamma}(-j)\boldsymbol{\Phi}_j' \equiv \boldsymbol{\Gamma}(0) - \boldsymbol{\Gamma}_{(p)}' \boldsymbol{\Phi}_{(p)} = \boldsymbol{\Gamma}(0) - \boldsymbol{\Phi}_{(p)}' \boldsymbol{\Gamma}_p \boldsymbol{\Phi}_{(p)}$$

In practical applications, these results can be used to derive *Yule--Walker estimates* of the parameters in the VAR(p) model by replacing the variance and covariance matrices by their estimates.

14.2.3 Special Case: VAR(1) Model

To examine the properties of VAR models in more detail, we will consider the VAR(1) model,

$$\boldsymbol{Z}_t = \boldsymbol{\Phi}\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t$$

where the mean vector $\boldsymbol{\mu}$ is assumed to be zero for convenience. For $k = 2$, we have the bivariate VAR(1) process

$$\boldsymbol{Z}_t = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \boldsymbol{Z}_{t-1} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}$$

or equivalently

$$\begin{aligned} z_{1t} &= \phi_{11} z_{1,t-1} + \phi_{12} z_{2,t-1} + a_{1t} \\ z_{2t} &= \phi_{21} z_{1,t-1} + \phi_{22} z_{2,t-1} + a_{2t} \end{aligned}$$

where ϕ_{11} and ϕ_{22} reflect the dependence of each component on its own past. The parameter ϕ_{12} shows the dependence of z_{1t} on $z_{2,t-1}$ in the presence of $z_{1,t-1}$, while ϕ_{21} shows the dependence of z_{2t} on $z_{1,t-1}$ in the presence of $z_{2,t-1}$. Thus, if $\phi_{12} \neq 0$ and $\phi_{21} \neq 0$, then there is a feedback relationship between the two components. On the other hand, if the off-diagonal elements of the parameter matrix Φ are zero, that is, $\phi_{12} = \phi_{21} = 0$, then z_{1t} and z_{2t} are not dynamically correlated. However, they are still contemporaneously correlated unless $\boldsymbol{\Sigma}$ is a diagonal matrix.

Relationship to Transfer Function Model. If $\phi_{12} = 0$, but $\phi_{21} \neq 0$, then z_{1t} does not depend on past values of z_{2t} but z_{2t} depends on past values of z_{1t}. A *transfer function relationship* then exists with z_{1t} acting as an input variable and z_{2t} as an output variable. However, unless z_{1t} is uncorrelated with a_{2t}, the resulting model is not in the standard transfer function form discussed in Chapter 12. To obtain the standard transfer function model, we let $a_{1t} = b_{1t}$ and $a_{2t} = \beta a_{1t} + b_{2t}$, where β is the regression coefficient of a_{2t} on a_{1t}. Under normality, the error term b_{2t} is then independent of a_{1t} and hence of b_{1t}. The unidirectional transfer function model is obtained by rewriting the equations for z_{1t} and z_{2t} above in terms of the orthogonal innovations b_{1t} and b_{2t}. This yields

$$(1 - \phi_{22}B)z_{2t} = \{\beta + (\phi_{21} - \beta\phi_{11})B\}z_{1,t-1} + b_{2t}$$

where the input variable z_{1t} does not depend on the noise term b_{2t}.

Hence, the bivariate transfer function model emerges as a special case of the bivariate AR model, in which a unidirectional relationship exists between the variables. In general, for a VAR(1) model in higher dimensions, $k > 2$, if the k series can be arranged so that the matrix $\mathbf{\Phi}$ is lower triangular, then the VAR(1) model can also be expressed in the form of unidirectional transfer function equations.

Stationarity Conditions for VAR(1) Model. The VAR(1) process is stationary if the roots of $\det\{\mathbf{I} - \mathbf{\Phi}B\} = 0$ exceed one in absolute value. Since $\det\{\mathbf{I} - \mathbf{\Phi}B\} = 0$ if and only if $\det\{\lambda\mathbf{I} - \mathbf{\Phi}\} = 0$ with $\lambda = 1/B$, it follows that the stationarity condition for the AR(1) model is equivalent to requiring that the eigenvalues of $\mathbf{\Phi}$ be less than one in absolute value. When this condition is met, the process has the convergent infinite MA representation (14.2.2) with MA coefficient matrices $\mathbf{\Psi}_j = \mathbf{\Phi}^j$, since from (14.2.3) the $\mathbf{\Psi}_j$ now satisfy

$$\mathbf{\Psi}_j = \mathbf{\Phi}\mathbf{\Psi}_{j-1} \equiv \mathbf{\Phi}^j\mathbf{\Psi}_0$$

To look at the stationarity for a k-dimensional VAR(1) model further, we note that for arbitrary $n > 0$, by $t + n$ successive substitutions in the right-hand side of $\boldsymbol{Z}_t = \mathbf{\Phi}\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t$ we obtain

$$\boldsymbol{Z}_t = \sum_{j=0}^{t+n} \mathbf{\Phi}^j \boldsymbol{a}_{t-j} + \mathbf{\Phi}^{t+n+1}\boldsymbol{Z}_{-n-1}$$

Hence, provided that all eigenvalues of $\mathbf{\Phi}$ are less than one in absolute value, as $n \to \infty$ this will converge to the infinite MA representation $\boldsymbol{Z}_t = \sum_{j=0}^{\infty} \mathbf{\Phi}^j \boldsymbol{a}_{t-j}$, with $\sum_{j=0}^{\infty} \|\mathbf{\Phi}^j\| < \infty$, which is stationary. For example, suppose that $\mathbf{\Phi}$ has k distinct eigenvalues $\lambda_1, \ldots, \lambda_k$, so there is a $k \times k$ nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{\Phi}\mathbf{P} = \mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_k)$. Then $\mathbf{\Phi} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$ and $\mathbf{\Phi}^j = \mathbf{P}\mathbf{\Lambda}^j\mathbf{P}^{-1}$, where $\mathbf{\Lambda}^j = \text{diag}(\lambda_1^j, \ldots, \lambda_k^j)$, so when all $|\lambda_i| < 1$, $\sum_{j=0}^{\infty} \|\mathbf{\Phi}^j\| < \infty$ since then $\sum_{j=0}^{\infty} \|\mathbf{\Lambda}^j\| < \infty$.

Moment Equations. For the VAR(1) model, the matrix Yule–Walker equations (14.2.5) simplify to

$$\mathbf{\Gamma}(l) = \mathbf{\Gamma}(l-1)\mathbf{\Phi}' \qquad \text{for } l \geq 1$$

so $\boldsymbol{\Gamma}(1) = \boldsymbol{\Gamma}(0)\boldsymbol{\Phi}'$, in particular, with

$$\boldsymbol{\Gamma}(0) = \boldsymbol{\Gamma}(-1)\boldsymbol{\Phi}' + \Sigma = \boldsymbol{\Phi}\boldsymbol{\Gamma}(0)\boldsymbol{\Phi}' + \boldsymbol{\Sigma}$$

Hence, $\boldsymbol{\Phi}'$ can be determined from $\boldsymbol{\Gamma}(0)$ and $\boldsymbol{\Gamma}(1)$ as $\boldsymbol{\Phi}' = \boldsymbol{\Gamma}(0)^{-1}\boldsymbol{\Gamma}(1)$ and also $\boldsymbol{\Gamma}(l) = \boldsymbol{\Gamma}(0)\boldsymbol{\Phi}'^{l}$. This last relation illustrates that the behavior of all correlations in $\boldsymbol{\rho}(l)$, obtained using (14.1.2), will be controlled by the behavior of the $\lambda_i^l, i = 1, \ldots, k$, where $\lambda_1, \ldots, \lambda_k$ are the eigenvalues of $\boldsymbol{\Phi}$, and shows that even the simple VAR(1) model is capable of fairly general correlation structures (e.g., mixtures of exponential decaying and damping sinusoidal terms) for dimensions $k > 1$. (For more details, see Reinsel, 1997, Section 2.2.3).

14.2.4 Numerical Example

Consider the bivariate ($k = 2$) AR(1) model $(\mathbf{I} - \boldsymbol{\Phi}B)\boldsymbol{Z}_t = \boldsymbol{a}_t$ with

$$\boldsymbol{\Phi} = \begin{bmatrix} 0.8 & 0.7 \\ -0.4 & 0.6 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$$

The roots of $\det\{\lambda\mathbf{I} - \boldsymbol{\Phi}\} = \lambda^2 - 1.4\lambda + 0.76 = 0$ are $\lambda = 0.7 \pm 0.5196i$, with absolute value equal to $(0.76)^{1/2}$; hence, the AR(1) model is stationary. Since the roots are complex, the correlations of this AR(1) process will exhibit damped sinusoidal behavior. The covariance matrix $\boldsymbol{\Gamma}(0)$ is determined by solving the linear equations $\boldsymbol{\Gamma}(0) - \boldsymbol{\Phi}\boldsymbol{\Gamma}(0)\boldsymbol{\Phi}' = \boldsymbol{\Sigma}$. Together with $\boldsymbol{\Gamma}(l) = \boldsymbol{\Gamma}(l-1)\boldsymbol{\Phi}'$, these lead to the covariance matrices

$$\boldsymbol{\Gamma}(0) = \begin{bmatrix} 18.536 & -1.500 \\ -1.500 & 8.884 \end{bmatrix} \qquad \boldsymbol{\Gamma}(1) = \begin{bmatrix} 13.779 & -8.315 \\ 5.019 & 5.931 \end{bmatrix}$$

$$\boldsymbol{\Gamma}(2) = \begin{bmatrix} 5.203 & -10.500 \\ 8.166 & 1.551 \end{bmatrix} \qquad \boldsymbol{\Gamma}(3) = \begin{bmatrix} -3.188 & -8.381 \\ 7.619 & -2.336 \end{bmatrix}$$

$$\boldsymbol{\Gamma}(4) = \begin{bmatrix} -8.417 & -3.754 \\ 4.460 & -4.449 \end{bmatrix} \qquad \boldsymbol{\Gamma}(5) = \begin{bmatrix} -9.361 & 1.115 \\ 0.453 & -4.453 \end{bmatrix}$$

The corresponding correlation matrices are obtained from $\boldsymbol{\rho}(l) = \mathbf{V}^{-1/2}\boldsymbol{\Gamma}(l)\mathbf{V}^{-1/2}$, where $\mathbf{V}^{-1/2} = \text{diag}(18.536^{-1/2}, 8.884^{-1/2})$. The autocorrelations and cross-correlations of this process are displayed up to 18 lags in Figure 14.1. We note that the correlation patterns are rather involved and correlations do not die out very quickly. The coefficients $\boldsymbol{\Psi}_j = \Phi^j$, $j \geq 1$, in the infinite MA representation for this AR(1) process are

$$\boldsymbol{\Psi}_1 = \begin{bmatrix} 0.80 & 0.70 \\ -0.40 & 0.60 \end{bmatrix} \qquad \boldsymbol{\Psi}_2 = \begin{bmatrix} 0.36 & 0.98 \\ -0.56 & 0.08 \end{bmatrix} \qquad \boldsymbol{\Psi}_3 = \begin{bmatrix} -0.10 & 0.84 \\ -0.48 & -0.34 \end{bmatrix}$$

$$\boldsymbol{\Psi}_4 = \begin{bmatrix} -0.42 & 0.43 \\ -0.25 & -0.54 \end{bmatrix} \qquad \boldsymbol{\Psi}_5 = \begin{bmatrix} -0.51 & -0.03 \\ 0.02 & -0.50 \end{bmatrix} \qquad \boldsymbol{\Psi}_6 = \begin{bmatrix} -0.39 & -0.38 \\ 0.22 & -0.28 \end{bmatrix}$$

So the elements of the $\boldsymbol{\Psi}_j$ matrices are also persistent and exhibit damped sinusoidal behavior similar to that of the correlations.

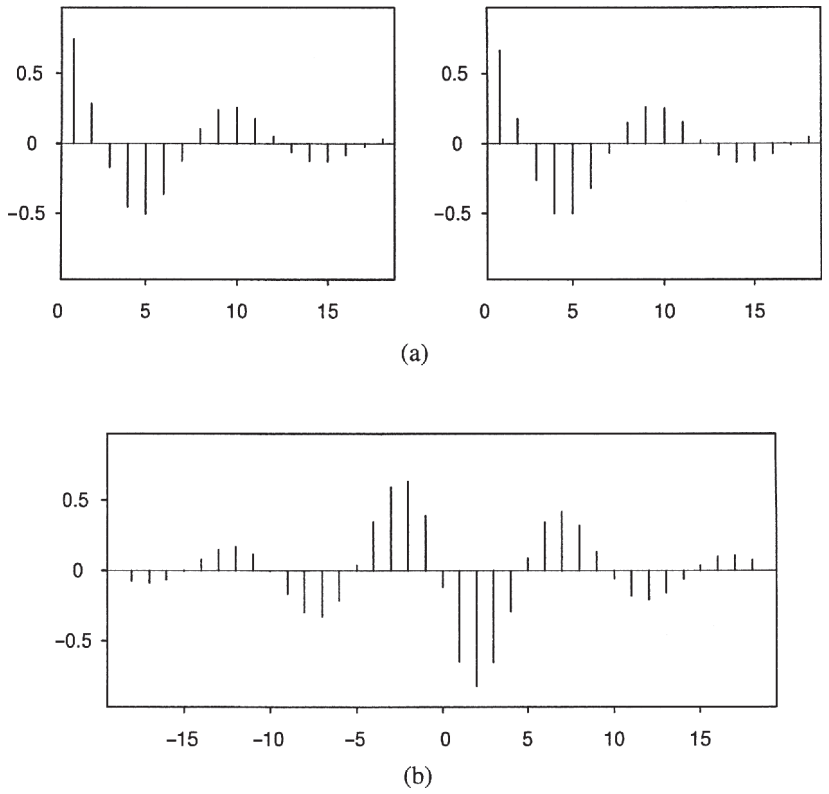

FIGURE 14.1 Theoretical autocorrelations and cross-correlations, $\rho_{ij}(l)$, for the bivariate VAR(1) process example: (a) autocorrelations $\rho_{11}(l)$ and $\rho_{22}(l)$ and (b) cross-correlations $\rho_{12}(l)$.

Finally, since $\det\{\lambda\mathbf{I} - \mathbf{\Phi}\} = \lambda^2 - 1.4\lambda + 0.76 = 0$, it follows from Reinsel (1997, Section 2.2.4) that each individual series z_{it} has a univariate ARMA(2, 1) model representation as $(1 - 1.4B + 0.76B^2)z_{it} = (1 - \eta_i B)\varepsilon_{it}$, $\sigma^2_{\varepsilon_i} = \text{var}[\varepsilon_{it}]$, where η_i and $\sigma^2_{\varepsilon_i}$ are readily determined. For a k-dimensional VAR(p) model, it can be shown that each individual component z_{it} follows a univariate ARMA of maximum order $(kp, (k-1)p)$. The order can be much less if the AR and MA polynomials have common factors (e.g., Wei, 2006, Chapter 16).

Computations in R. The covariance matrices $\mathbf{\Gamma}(l)$ and the $\mathbf{\Psi}$ matrices shown above can be reproduced using the MTS package in R as follows:

```
> library(MTS)
> phi1=matrix(c(0.8,-0.4,0.7,0.6),2,2)
> sig=matrix(c(4,1,1,2),2,2)
> eigen(phi1)
> m1=VARMAcov(Phi=phi1, Sigma=sig, lag=5)
> names(m1)
  [1] "autocov" "ccm"
> autocov=t(m1$autocov)
```

```
> m2=PSIwgt(Phi=phi1)
> names(m2)
  [1] "psi.weight" "irf"
> m2$psi.weight
```

The command VARMAcov() computes the covariance and cross-correlation matrices up to 12 lags by default. These matrices need to be transposed using the command t() since MTS defines the lag l covariance matrix $\boldsymbol{\Gamma}(l)$ as $E[(\boldsymbol{Z}_t - \boldsymbol{\mu})(\boldsymbol{Z}_{t-l} - \boldsymbol{\mu})']$, whereas the definition $E[(\boldsymbol{Z}_{t-l} - \boldsymbol{\mu})(\boldsymbol{Z}_t - \boldsymbol{\mu})']$ is used in this chapter. Transposing the matrices makes the results from R consistent with our definition. The command eigen(phi1) included in the code gives the eigenvalues of the matrix $\boldsymbol{\Phi}$.

14.2.5 Initial Model Building and Least-Squares Estimation for VAR Models

Given an observed vector time series $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N$ of length N from a multivariate process, the development of an appropriate VAR model for the series can be performed iteratively using a three-stage procedure of model specification, parameter estimation, and diagnostic checking. In the VAR case, the model specification involves choosing a suitable value for the order p. Some useful tools at this stage include the sample covariance and correlation matrices described below and the sample partial autoregression matrices discussed, for example, by Tiao and Box (1981). The latter quantities are analogous to the partial autocorrelations used in the univariate case and are estimated as the last autoregressive matrix, $\boldsymbol{\Phi}_m$, in a VAR(m) model with $m = 1, 2, \ldots$. The estimates $\hat{\boldsymbol{\Phi}}_m$ can be derived from the Yule–Walker equations or by least-squares estimation of the parameter matrices. Statistical tests are used to determine the significance of the estimates for each value of m. The partial autoregression matrices are zero for all lags greater than p and are thus particularly useful for identifying the autoregressive model order. Additional methods for model selection include the use of information criteria such as AIC, BIC, and HC, as well as methods based on canonical correlation analysis described later in this chapter.

Sample Covariance and Correlation Matrices. Given an observed time series, the sample covariance matrix of the $\boldsymbol{Z}_t$ at lag l is defined as

$$\hat{\boldsymbol{\Gamma}}(l) = \mathbf{C}(l) = \frac{1}{N}\sum_{t=1}^{N-l}(\boldsymbol{Z}_t - \overline{\boldsymbol{Z}})(\boldsymbol{Z}_{t+l} - \overline{\boldsymbol{Z}})' \quad l = 0, 1, 2, \ldots \tag{14.2.6}$$

where $\overline{\boldsymbol{Z}} = (\overline{z}_1, \ldots, \overline{z}_k)' = N^{-1}\sum_{t=1}^{N}\boldsymbol{Z}_t$ is the sample mean vector, which is a natural estimator of the process mean vector $\boldsymbol{\mu} = E[\boldsymbol{Z}_t]$ in the stationary case. In particular, $\hat{\boldsymbol{\Gamma}}(0) = \mathbf{C}(0) = N^{-1}\sum_{t=1}^{N}(\boldsymbol{Z}_t - \overline{\boldsymbol{Z}})(\boldsymbol{Z}_t - \overline{\boldsymbol{Z}})'$ is the sample covariance matrix of the $\boldsymbol{Z}_t$. The (i, j)th element of $\hat{\boldsymbol{\Gamma}}(l)$ is given by

$$\hat{\gamma}_{ij}(l) = c_{ij}(l) = \frac{1}{N}\sum_{t=1}^{N-l}(z_{it} - \overline{z}_i)(z_{j,t+l} - \overline{z}_j)$$

The *sample cross-correlations* are defined as

$$\hat{\rho}_{ij}(l) = r_{ij}(l) = \frac{c_{ij}(l)}{\{c_{ii}(0)c_{jj}(0)\}^{1/2}} \quad i, j = 1, \ldots, k$$

For a stationary series, the $\hat{\rho}_{ij}(l)$ are sample estimates of the theoretical $\rho_{ij}(l)$. The asymptotic sampling properties of sample correlations $\hat{\rho}_{ij}(l)$ were discussed earlier in Section 12.1.3. The expressions for the asymptotic variances and covariances of the estimates are complicated but simplify in certain cases. For example, in the special case where $\boldsymbol{Z}_t$ is a white noise process, the results give $\text{var}[\hat{\rho}_{ij}(l)] \simeq 1/(N-l)$.

The sample cross-correlation matrices are important tools for the initial specification of a model for the series $\boldsymbol{Z}_t$. They are particularly useful in the model specification for a low-order pure vector moving average model, which has the property that $\rho_{ij}(l) = 0$ for all $l > q$, as discussed in Section 14.3 below. However, similar to the univariate case, a slowly decaying pattern in the estimated autocorrelation and cross-correlation matrices would indicate that autoregressive terms are needed.

Estimation of the Partial Autoregression Matrices. Consider the vector autoregressive model of order m, $\boldsymbol{Z}_t = \boldsymbol{\delta} + \sum_{j=1}^{m} \boldsymbol{\Phi}_j \boldsymbol{Z}_{t-j} + \boldsymbol{a}_t$, where $\boldsymbol{\delta} = (\mathbf{1} - \boldsymbol{\Phi}_1 - \cdots - \boldsymbol{\Phi}_m)\boldsymbol{\mu}$ accounts for the non-zero mean vector. Estimates of the partial autoregressive matrices can be obtained from the Yule–Walker equations in (14.2.5) as

$$\hat{\boldsymbol{\Phi}}_{(m)} = [\hat{\boldsymbol{\Phi}}_{1m}, \ldots, \hat{\boldsymbol{\Phi}}_{mm}]' = \hat{\boldsymbol{\Gamma}}_m^{-1} \hat{\boldsymbol{\Gamma}}_{(m)}$$

The estimate of the error covariance matrix estimate is $\hat{\boldsymbol{\Sigma}}_m = \hat{\boldsymbol{\Gamma}}(0) - \sum_{j=1}^{m} \hat{\boldsymbol{\Gamma}}(-j)\hat{\boldsymbol{\Phi}}_{jm}'$. The estimation is performed for $m = 1, 2, \ldots$, yielding a sequence of estimates $\hat{\boldsymbol{\Phi}}_{mm}$ of the last parameter matrix in the VAR(m) model. These matrices are referred to as partial autoregression matrices by Tiao and Box (1981).

An asymptotically equivalent procedure is to estimate the partial autoregression matrices using multivariate linear least-squares (LS) estimation described, for example, by Johnson and Wichern (2007). Using this approach, the components of $\boldsymbol{Z}_t$ are regressed on the lagged vector values $\boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_{t-m}$, by first writing the VAR(m) model in regression form as

$$\boldsymbol{Z}_t = \boldsymbol{\delta} + \sum_{j=1}^{m} \boldsymbol{\Phi}_j \boldsymbol{Z}_{t-j} + \boldsymbol{a}_t = \boldsymbol{\delta} + \boldsymbol{\Phi}_{(m)}' \boldsymbol{X}_t + \boldsymbol{a}_t \tag{14.2.7}$$

with $\boldsymbol{X}_t = (\boldsymbol{Z}_{t-1}', \ldots, \boldsymbol{Z}_{t-m}')'$. The LS estimates for the AR parameters are then given by

$$\hat{\boldsymbol{\Phi}}_{(m)} = [\hat{\boldsymbol{\Phi}}_{1m}, \ldots, \hat{\boldsymbol{\Phi}}_{mm}]' = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\boldsymbol{Z}} \tag{14.2.8}$$

where the matrices $\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{X}}$, respectively, have typical rows $(\boldsymbol{Z}_t - \overline{\boldsymbol{Z}}_{(0)})'$ and

$$[(\boldsymbol{Z}_{t-1} - \overline{\boldsymbol{Z}}_{(1)})', \ldots, (\boldsymbol{Z}_{t-m} - \overline{\boldsymbol{Z}}_{(m)})'] \quad t = m+1, \ldots, N$$

with $\overline{\boldsymbol{Z}}_{(i)} = n^{-1} \sum_{t=m+1}^{N} \boldsymbol{Z}_{t-i}$ and $n = N - m$. The estimate of the error covariance matrix $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}}_m = [n - (km+1)]^{-1} \mathbf{S}_m \tag{14.2.9}$$

where

$$\mathbf{S}_m = \sum_{t=m+1}^{N} \hat{\boldsymbol{a}}_t \hat{\boldsymbol{a}}_t'$$

is the residual sum-of-squares matrix and

$$\hat{\boldsymbol{a}}_t = (\boldsymbol{Z}_t - \overline{\boldsymbol{Z}}_{(0)}) - \sum_{j=1}^{m} \hat{\boldsymbol{\Phi}}_j (\boldsymbol{Z}_{t-j} - \overline{\boldsymbol{Z}}_{(j)})$$

are the residual vectors. These LS estimators $\hat{\boldsymbol{\Phi}}_j$ are also the conditional maximum likelihood (ML) estimators under the normality assumption. Asymptotic distribution theory for the LS estimators in the stationary VAR model was provided by Hannan (1970, Chapter 6). Under a stationary VAR(m) model, the distribution of vec[$\hat{\boldsymbol{\Phi}}_{(m)}$] is approximately multivariate normal with mean vector vec[$\boldsymbol{\Phi}_{(m)}$] and covariance matrix estimated by $\hat{\boldsymbol{\Sigma}}_m \otimes (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$, where $\otimes$ denotes the Kronecker product of $\hat{\boldsymbol{\Sigma}}_m$ and $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$.

Sequential Likelihood Ratio Tests. The estimation of the partial autoregression matrices is supplemented by likelihood ratio tests that are applied sequentially to help determine the model order p. (e.g., see Tiao and Box (1981) and Reinsel (1997, Chapter 4)). Thus, after fitting a VAR(m) model, we test the null hypothesis H_0: $\boldsymbol{\Phi}_{mm} = \mathbf{0}$ against the alternative $\boldsymbol{\Phi}_{mm} \neq \mathbf{0}$, using the likelihood ratio (LR) statistic

$$M_m = -\left(n - mk - \frac{1}{2}\right) \ln\left[\frac{|\mathbf{S}_m|}{|\mathbf{S}_{m-1}|}\right] \tag{14.2.10}$$

where $\mathbf{S}_m$ is the residual sum-of-squares matrix defined above, and $n = N - m - 1$ is the effective number of observations assuming that the model includes a constant term. For large n, when H_0: $\boldsymbol{\Phi}_{mm} = \mathbf{0}$ is true, the statistic M_m has an approximate χ^2 distribution with k^2 degrees of freedom, and we reject H_0 for large values of M_m. The LR test statistic in (14.2.10) is asymptotically equivalent to a Wald statistic formed in terms of the LS estimator $\hat{\boldsymbol{\Phi}}_{mm}$ of $\boldsymbol{\Phi}_{mm}$.

This procedure is a natural extension of the use of the sample PACF $\hat{\phi}_{mm}$ for identification of the order of an AR model in the univariate case as described in Section 6.2. However, unlike the univariate case, the partial autoregression matrices are not partial autocorrelation matrices (or correlations of any kind) in the vector case. Similar tests based on the sample partial autocorrelation matrices, whose elements are proper correlation coefficients, are described by Reinsel (1997, Chapter 4) and Wei (2006, Chapter 16).

Use of Information Criteria. Model selection criteria such as AIC, BIC, and HQ can also be employed for model specification. Here, AIC represents Akaike's information criterion (Akaike, 1974a), BIC is the Bayesian information criterion due to Schwarz (1978), and HQ is the model selection criterion proposed by Hannan and Quinn (1979); see also Quinn (1980). These criteria are likelihood based and include under normality the determinant of the innovations covariance matrix that reflects the goodness of fit of the model. A second term is a function of the number of fitted parameters and penalizes models that are unnecessarily complex. For the VAR model, we have

$$\begin{aligned}
\text{AIC}_m &= \ln\{|\tilde{\boldsymbol{\Sigma}}_m|\} + 2mk^2/N \\
\text{BIC}_m &= \ln\{|\tilde{\boldsymbol{\Sigma}}_m|\} + mk^2 \ln(N)/N \\
\text{HQ}_m &= \ln\{|\tilde{\boldsymbol{\Sigma}}_m|\} + 2mk^2 \ln(\ln(N))/N
\end{aligned}$$

where N is the sample size, m is the VAR order, and $\tilde{\boldsymbol{\Sigma}}_m$ is the corresponding ML residual covariance matrix estimate of $\boldsymbol{\Sigma}$. It can be seen that BIC imposes a greater ''penalty factor'' for the number of estimated parameters than does AIC, while HQ is intermediate between AIC and BIC. Other similar measures include the final prediction error (FPE) criterion suggested by Akaike (1971). These criteria can be used to compare models fitted using maximum likelihood and the model that gives the lowest value for a given criterion would be selected. For a discussion of the properties and performance of different model selection criteria, see, for example, Quinn (1980) and Lütkepohl (2006).

14.2.6 Parameter Estimation and Model Checking

Parameter Estimation. With the order of the VAR model specified, the model parameters can be estimated using the least-squares procedure described above. For a stationary process, the Yule–Walker estimates are asymptotically equivalent to the least-squares estimates. However, when the process is nonstationary or near nonstationary, it is known that the least-squares estimator still performs consistently, whereas the Yule–Walker estimator may have a considerable bias. Hence, the least-squares method is generally to be preferred (e.g., Reinsel, 1997, Section 4.4). Under the normality assumption, the least-squares estimates are equivalent to conditional maximum likelihood estimates. Exact maximum likelihood estimates can be derived using the unconditional likelihood function described for VARMA models in Section 14.4.5. However, use of the conditional likelihood function simplifies the calculations and is often adequate for VAR models in practice.

Model Checking. Model diagnostics of the estimated VAR model are primarily based on examination of the residual vectors $\hat{\boldsymbol{a}}_t$ from the estimated model and their sample covariance matrices. The residuals $\hat{\boldsymbol{a}}_t$ are calculated from (14.2.1) with the parameters replaced by their estimates $\hat{\boldsymbol{\Phi}}_j$. Useful diagnostic checks include plots of the residuals against time and/or against other variables, and detailed examination of the cross-correlation matrices of the residuals. Approximate two-standard-error limits can be imposed to assess the statistical significance of the residual correlations.

In addition, overall portmanteau or ''goodness-of-fit'' tests based on the residual covariance matrices at several lags can be employed for model checking; see, for example, Hosking (1980), Li and McLeod (1981), Poskitt and Tremayne (1982), and Ali (1989). Specifically, using s lags, an overall goodness-of-fit test statistic, analogous to that proposed by Ljung and Box (1978) for the univariate case, is given by

$$Q_s = N^2 \sum_{l=1}^{s} (N-l)^{-1} \mathrm{tr}[\hat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{a}}}(l) \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{a}}}(l)' \hat{\boldsymbol{\Sigma}}^{-1}] \tag{14.2.11}$$

where

$$\hat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{a}}}(l) = N^{-1} \sum_{t=1}^{N-l} \hat{\boldsymbol{a}}_t \hat{\boldsymbol{a}}'_{t+l} \quad l = 0, 1, \ldots, s$$

with $\hat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{a}}}(0) \approx \hat{\boldsymbol{\Sigma}}$. Under the null hypothesis of model adequacy, the test statistic Q_s is approximately distributed as chi-squared with $k^2(s-p)$ degrees of freedom. The fitted model is rejected as inadequate for large values of Q_s. Mahdi and McLeod (2012) extended the portmanteau test of Peña and Rodríguez (2002, 2006) described in Chapter 8 to the

multivariate case and proposed a test based on the determinant of the autocorrelation matrix of the multivariate residuals. Alternative tests such as score or Lagrange multiplier (LM) tests have also been proposed in the literature. For a discussion of the LM tests and their relationship to portmanteau tests, see, for example, Reinsel (1997) and Lütkepohl (2006).

14.2.7 An Empirical Example

To illustrate the model building procedure for a vector process outlined above, we consider the bivariate time series of U.S. fixed investment and change in business inventories. These data are quarterly, seasonally adjusted, and are given in Lütkepohl (2006). The fixed investment data for the time period 1947 to 1971 are shown in Figure 14.2, and the changes in business inventories series for the same period are shown in Figure 14.3(b). Since the investment series is clearly nonstationary, the first differences of this series, which are displayed in Figure 14.3(a), are considered as series z_{1t} together with the change in business inventories as series z_{2t}, resulting in $N = 99$ quarterly observations.

Sample cross-correlation matrices of the series $\boldsymbol{Z}_t = (z_{1t}, z_{2t})'$ for lags 1 through 12 are shown in Table 14.1, and these sample autocorrelations and cross-correlations $\hat{\rho}_{ij}(l)$ are also displayed up to 18 lags in Figure 14.4. Included in Figure 14.4 are the rough guidelines of the two-standard-error limits $\pm 2/\sqrt{N} \simeq \pm 0.2$, which are appropriate for the $\hat{\rho}_{ij}(l)$ from a vector white noise process as noted in Section 14.2.5. These sample correlations show exponentially decaying and damped sinusoidal behavior as a function of lag l, indicative of autoregressive dependence structure in the series.

To select a suitable model, we apply the sequential likelihood ratio test and the three information criteria discussed above to the data. The calculations are performed using the MTS package in R and the results are summarized in Table 14.2. We note that the three criteria AIC_m, BIC_m, and HQ_m all attain a minimum at $m = 2$. The likelihood ratio statistic M_m also supports the value $m = 2$, although a slight discrepancy occurs at $m = 4$. These results therefore indicate that, among pure autoregressive models, a second-order VAR(2) model may be the most appropriate for these data.

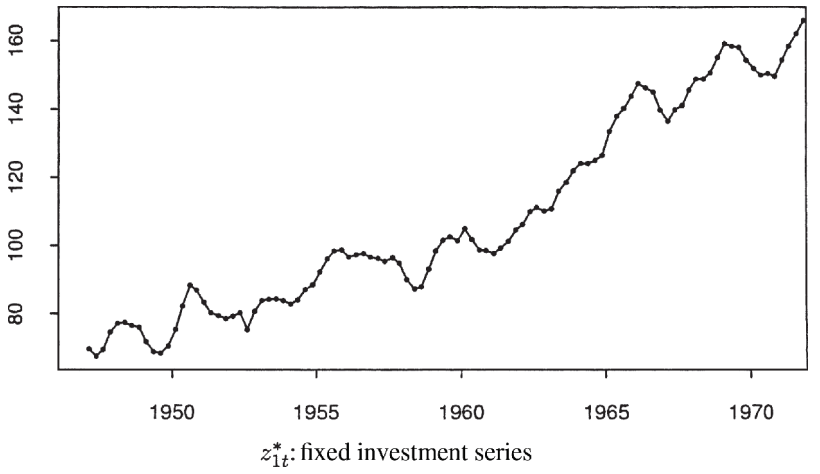

FIGURE 14.2 Quarterly (seasonally adjusted) U.S. fixed investment data for 1947 through 1971.

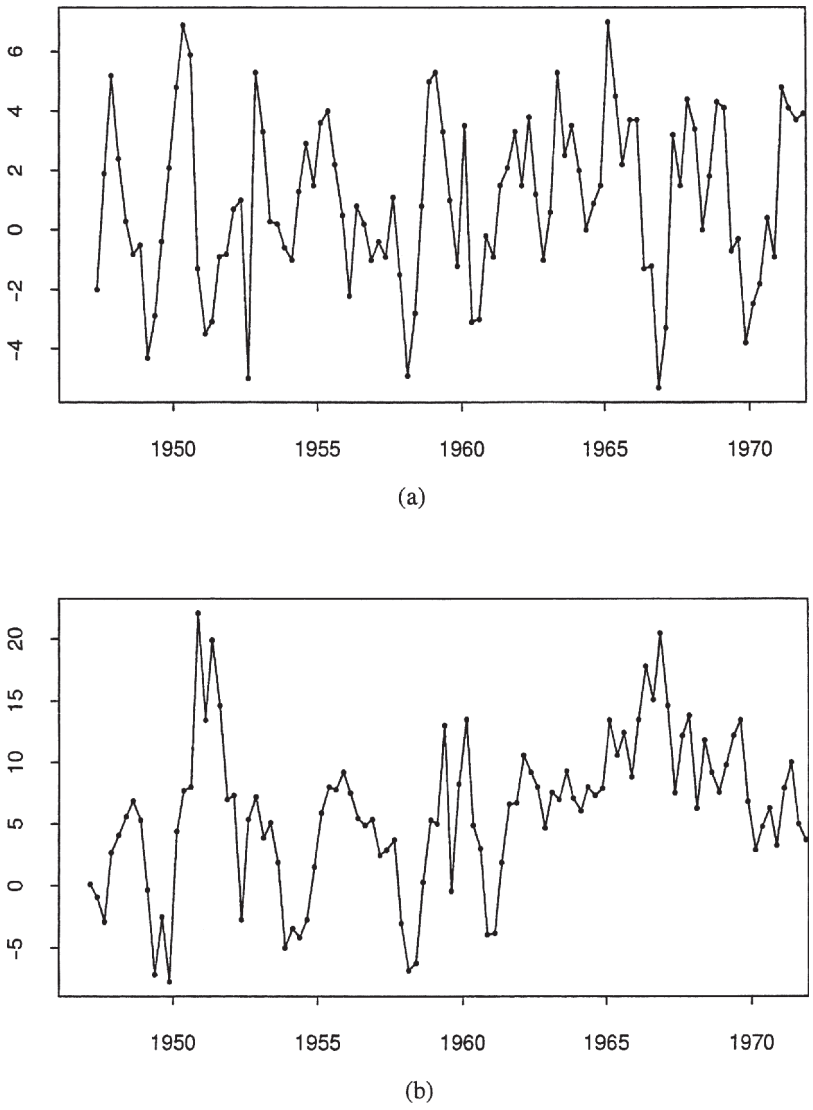

FIGURE 14.3 Quarterly (seasonally adjusted) first differences of U.S. fixed investment data and changes in business inventories data (in billions) for the period 1947 through 1971: (a) z_{1t}: first differences of investment series, $z_{1t} = z_{1t}^* - z_{1,t-1}^*$; and (b) z_{2t}: changes in business inventories series.

TABLE 14.1 Sample Correlation Matrices $\hat{\rho}(l)$ for the Bivariate Quarterly Series of First Differences of U.S. Fixed Investment and U.S. Changes in Business Inventories

l	1		2		3		4		5		6	
$\hat{\rho}(l)$	0.47	0.27	0.10	0.35	−0.12	0.29	−0.31	0.27	−0.30	0.19	−0.21	0.04
	−0.06	0.68	−0.33	0.50	−0.29	0.32	−0.21	0.07	−0.10	0.07	0.10	0.01
l	7		8		9		10		11		12	
$\hat{\rho}(l)$	−0.14	−0.04	−0.09	−0.11	0.13	−0.03	0.19	0.07	0.13	0.12	0.02	0.20
	0.15	0.04	0.20	0.05	0.12	0.05	0.05	0.11	0.01	0.09	−0.04	0.05

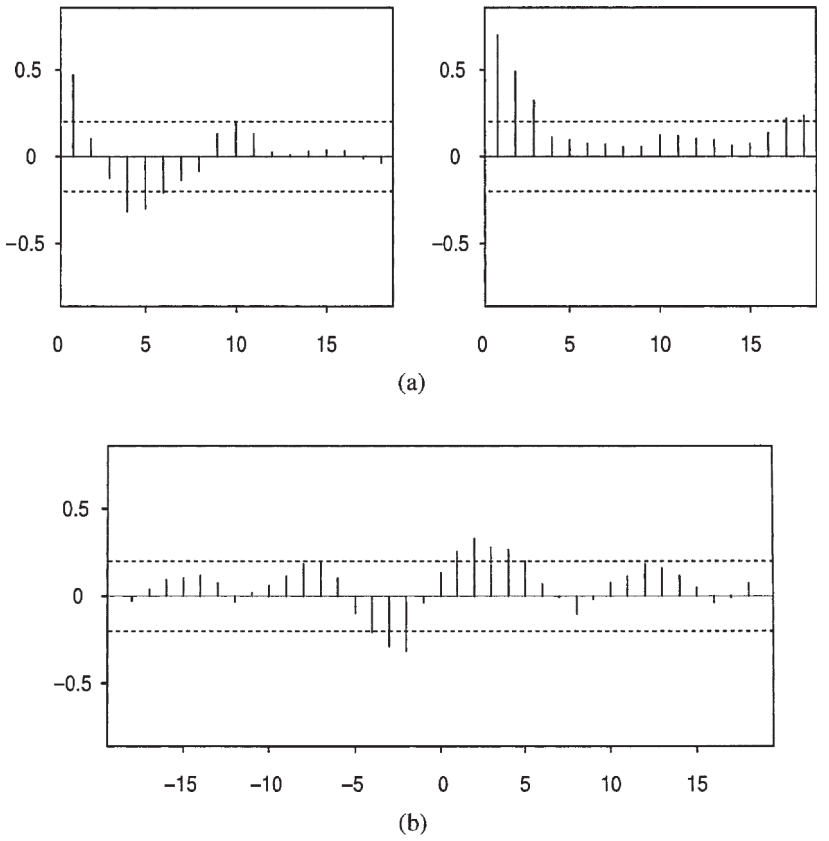

FIGURE 14.4 Sample auto- and cross-correlations $\hat{\rho}_{ij}(l)$ for the bivariate series of first differences of U.S. fixed investment and U.S. changes in business inventories: (a) sample autocorrelations $\hat{\rho}_{11}(l)$ and $\hat{\rho}_{22}(l)$ and (b) sample cross-correlations $\hat{\rho}_{12}(l)$.

TABLE 14.2 Order Selection Statistics for the U.S. Business Investment and Inventories Data

m (VAR Order)	AIC_m	BIC_m	HQ_m	M_m	p-Value
0	5.539	5.539	5.539	0.000	0.000
1	4.723	4.828	4.766	73.997	0.000
2	4.597	4.807	4.682	16.652	0.002
3	4.659	4.974	4.786	1.483	0.830
4	4.614	5.033	4.784	9.628	0.047
5	4.624	5.148	4.836	5.283	0.260
6	4.703	5.332	4.958	0.113	0.999
7	4.759	5.493	5.056	1.785	0.775
8	4.785	5.623	5.124	3.755	0.440

The LS estimates from the AR(2) model (with estimated standard errors in parentheses), as well as the ML estimate of $\boldsymbol{\Sigma}$, are given as

$$\hat{\boldsymbol{\Phi}}_1 = \begin{bmatrix} 0.504 & 0.108 \\ (0.096) & (0.056) \\ 0.345 & 0.531 \\ (0.177) & (0.103) \end{bmatrix} \quad \hat{\boldsymbol{\Phi}}_2 = \begin{bmatrix} -0.146 & -0.205 \\ (0.099) & (0.054) \\ 0.256 & 0.139 \\ (0.181) & (0.099) \end{bmatrix}$$

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} 5.0270 & 1.6958 \\ 1.6958 & 16.9444 \end{bmatrix}$$

with $|\tilde{\boldsymbol{\Sigma}}| = 82.3032$. The estimates of the two constant terms are 1.217 and 1.527, with respective standard errors of 0.354 and 0.650. In the matrix $\hat{\boldsymbol{\Phi}}_2$, the coefficient estimate in the (1, 2) position is statistically significant, while the rest are insignificant and might perhaps be omitted.

We now examine the residuals $\hat{\boldsymbol{a}}_t$ from the fitted VAR(2) model. The residual autocorrelations and cross-correlations are displayed in Figure 14.5. The approximate two-standard-error limits are also included in the graphs. The individual elements of the residual correlation matrices are generally quite small for all lags through $l = 12$, with $|\hat{\rho}_{\hat{a},ij}(l)| \ll 2/\sqrt{N} = 0.2$ in nearly all cases. One notable feature of these residual correlations, however, is the (marginally) significant correlation of $\hat{\rho}_{\hat{a},22}(4) = -0.20$ at lag 4 for the second residual series $\hat{a}_{2t}$ (see lower right panel of Figure 14.5). This feature, which also appears visible from the p-values of the portmanteau test shown in Figure 14.6, may be a consequence of the seasonal adjustment procedure, related to a weak seasonal structure that may still exist in the quarterly (''seasonally adjusted'') series $\boldsymbol{Z}_t$. To accommodate this feature, we could consider a modification to the VAR(2) model by inclusion of an MA coefficient matrix $\boldsymbol{\Theta}_4$ at the quarterly seasonal lag of 4 in the model. Although this could lead to a small improvement, we do not pursue this modification here.

As a benchmark for comparison against the bivariate AR(2) model fitted above, comparable *univariate* models for z_{1t} and z_{2t} that were found to be adequate, estimated by the conditional ML method, were obtained as

$$(1 - 1.275B + 0.545B^2)z_{1t} = 0.251 + (1 - 0.769B)\varepsilon_{1t}$$

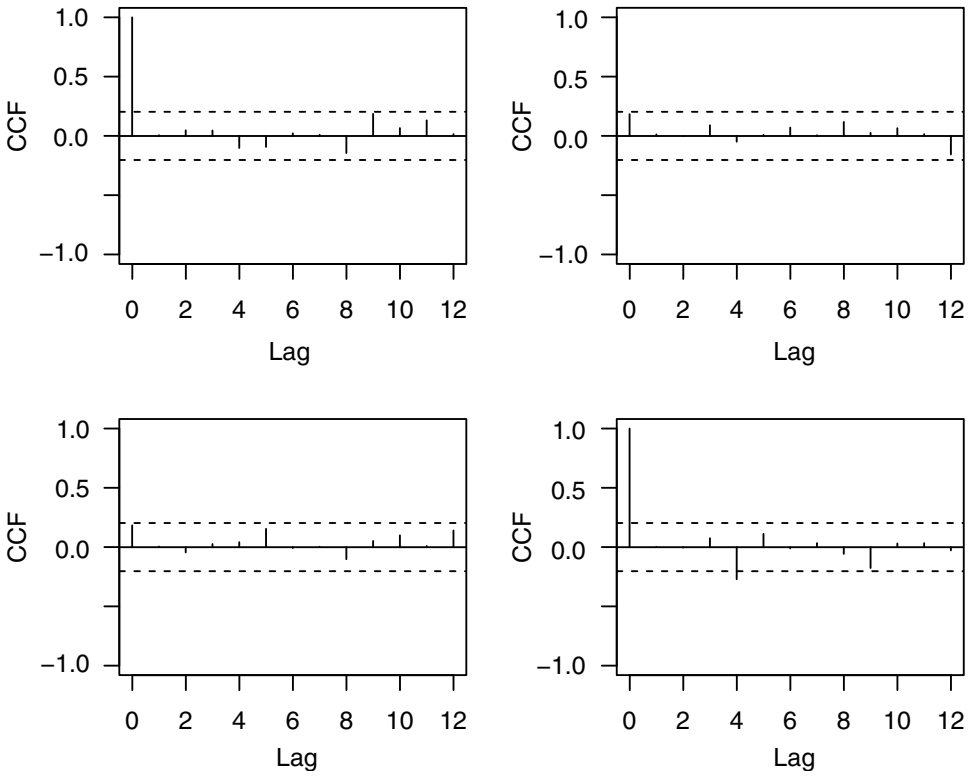

FIGURE 14.5 Cross-correlation matrices for the residuals from the VAR(2) model fitted to the U.S. business investment and inventories data.

with $\hat{\sigma}^2_{\varepsilon_1} = 5.44$, and $(1 - 0.690B)z_{2t} = 1.808 + \varepsilon_{2t}$, with $\hat{\sigma}^2_{\varepsilon_2} = 19.06$. Note that the residual variances are slightly larger in this case. The fitted bivariate models imply that the changes in business inventories series z_{2t} have a modest but significant influence on the (first differences of) investments z_{1t}, but there appears to be less influence in the feedback from investments to the changes in inventories series. In addition, there is only a small degree of contemporaneous correlation suggested, since the correlation between the residual series $\hat{a}_{1t}$ and $\hat{a}_{2t}$ in the bivariate models estimated from $\tilde{\boldsymbol{\Sigma}}$ equals 0.184.

Remark. The bivariate analysis described above was performed using the multivariate time series package MTS in R. Letting zz denote the data after differencing the investments series, the relevant commands are

```
> ccm(zz)            % Cross-correlation analysis
> m1=VARorder(zz)    % Order selection
> m2=VAR(zz,2)       % Estimation of VAR(2) model
> MTSdiag(m2)        % Model checking
> ccm(m2$residuals) % Residual cross-correlation analysis
```

For more detailed discussion and for demonstrations of the analysis capabilities of the MTS package in R, see Tsay (2014). Multivariate time series tools are also available in other

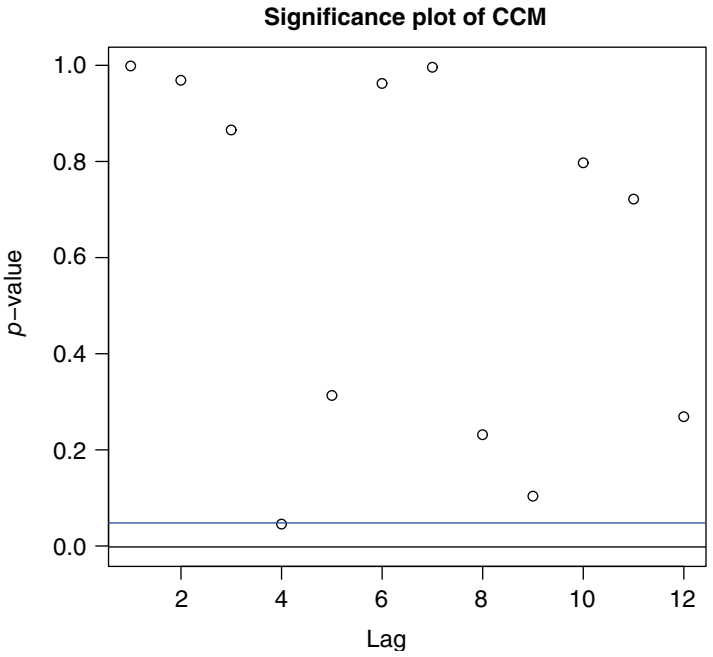

FIGURE 14.6 Plot of p-values of the multivariate portmanteau statistic applied to the residuals from the fitted VAR(2) model.

packages such as the SCA package released by Scientific Computing Associates Corp., and the S-Plus software package available from TIBCO Software, Inc.

14.3 VECTOR MOVING AVERAGE MODELS

The vector autoregressive models described above provide an adequate representation to many applied time series and are widely used in practice. However, pure autoregressive models have a disadvantage in that the model order needed to obtain a satisfactory representation can in some cases be rather high. Analogous to the univariate case, a more parsimonious representation can sometimes be achieved by adding moving average terms to the model. This would result in the vector ARMA (or VARMA) model form mentioned briefly in Section 14.1.4. Aggregation of vector series across time or in space also creates a need for VARMA models as noted e.g. by Lütkepohl and Poskitt (1996). In addition, trend or seasonal adjustments may change the dependence structure and make a pure VAR model inadequate (e.g., Maravall, 1993). Prior to discussing the VARMA model in more detail, we will briefly examine the special case when no autoregressive terms are present and the series follows a pure moving average model.

14.3.1 Vector MA(q) Model

A vector moving average model of order q, or VMA(q) model, is defined as

$$\boldsymbol{Z}_t = \boldsymbol{\mu} + \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j} \tag{14.3.1}$$

or equivalently, $\boldsymbol{Z}_t = \boldsymbol{\mu} + \boldsymbol{\Theta}(B)\boldsymbol{a}_t$, where $\boldsymbol{\mu}$ is the mean of the process, $\boldsymbol{\Theta}(B) = \mathbf{I} - \boldsymbol{\Theta}_1 B - \cdots - \boldsymbol{\Theta}_q B^q$ is a matrix polynomial of order q, and the $\boldsymbol{\Theta}_i$ are $k \times k$ matrices with $\boldsymbol{\Theta}_q \neq 0$.

Invertibility. A vector MA(q) process is said to be *invertible* if it can be represented in the form

$$(\boldsymbol{Z}_t - \boldsymbol{\mu}) - \sum_{j=1}^{\infty} \boldsymbol{\Pi}_j (\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}) = \boldsymbol{a}_t \tag{14.3.2}$$

or equivalently as $\boldsymbol{\Pi}(B)(\boldsymbol{Z}_t - \boldsymbol{\mu}) = \boldsymbol{a}_t$ where $\boldsymbol{\Pi}(B) = \mathbf{I} - \sum_{j=1}^{\infty} \boldsymbol{\Pi}_j B^j$, with $\sum_{j=1}^{\infty} \|\boldsymbol{\Pi}_j\| < \infty$. The process is invertible if all the roots of $\det\{\boldsymbol{\Theta}(B)\} = 0$ are greater than one in absolute value. The process then has the infinite VAR representation given by (14.3.2) with $\boldsymbol{\Pi}(B) = \boldsymbol{\Theta}^{-1}(B)$ so that $\boldsymbol{\Theta}(B)\boldsymbol{\Pi}(B) = \mathbf{I}$. As in the univariate case, this form is particularly useful for determining how forecasts of future observations depend on current and past values of the k series.

Moment Equations. For the VMA(q) model, the covariance matrices $\boldsymbol{\Gamma}(l)$ are given by

$$\boldsymbol{\Gamma}(l) = \sum_{h=0}^{q-l} \boldsymbol{\Theta}_h \boldsymbol{\Sigma} \boldsymbol{\Theta}'_{h+l} \tag{14.3.3}$$

for $l = 0, 1, \ldots, q$, with $\boldsymbol{\Theta}_0 = -\mathbf{I}$, and $\boldsymbol{\Gamma}(l) = \mathbf{0}$, for $l > q$. The result is readily verified since the $\{\boldsymbol{a}_t\}$ form a white noise sequence and $\mathrm{Cov}[\boldsymbol{\Theta}\boldsymbol{a}_t] = \boldsymbol{\Theta}\boldsymbol{\Sigma}\boldsymbol{\Theta}'$.

14.3.2 Special Case: Vector MA(1) Model

To examine the properties further, we consider the VMA(1) model, $\boldsymbol{Z}_t - \boldsymbol{\mu} = \boldsymbol{a}_t - \boldsymbol{\Theta}\boldsymbol{a}_{t-1}$. From the same reasoning as given concerning the stationarity condition for the VAR(1) process, the invertibility condition for the VMA(1) model is equivalent to all eigenvalues of $\boldsymbol{\Theta}$ being less than one in absolute value. Then we have the convergent infinite VAR representation (14.3.2) with infinite VAR coefficient matrices $\boldsymbol{\Pi}_j = -\boldsymbol{\Theta}^j$, $j \geq 1$. This follows since $\boldsymbol{\Theta}(B)\boldsymbol{\Pi}(B) = \mathbf{I}$ now simplifies to $\boldsymbol{\Pi}_j = \boldsymbol{\Theta}\boldsymbol{\Pi}_{j-l} \equiv \boldsymbol{\Theta}^j \boldsymbol{\Pi}_0$ with $\boldsymbol{\Pi}_0 = -\mathbf{I}$. Also, from (14.3.3) the covariance matrices of the VMA(1) process simplify to

$$\boldsymbol{\Gamma}(0) = \boldsymbol{\Sigma} + \boldsymbol{\Theta}\boldsymbol{\Sigma}\boldsymbol{\Theta}', \quad \boldsymbol{\Gamma}(1) = -\boldsymbol{\Sigma}\boldsymbol{\Theta}' = \boldsymbol{\Gamma}(-1)'$$

and $\boldsymbol{\Gamma}(l) = \mathbf{0}$ for $|l| > 1$. Thus, as in the univariate MA(1) case, all covariances are zero for lags greater than one.

14.3.3 Numerical Example

Consider the bivariate ($k = 2$) VMA(1) model $\boldsymbol{Z}_t = (\mathbf{I} - \boldsymbol{\Theta}B)\boldsymbol{a}_t$ with

$$\boldsymbol{\Theta} = \begin{bmatrix} 0.8 & 0.7 \\ -0.4 & 0.6 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$$

Similar to results for the VAR(1) example, the roots of $\det\{\lambda\mathbf{I} - \boldsymbol{\Theta}\} = \lambda^2 - 1.4\lambda + 0.76 = 0$ are $\lambda = 0.7 \pm 0.5196i$, with absolute value equal to $(0.76)^{1/2}$; hence, the VMA(1) model is invertible. The coefficient matrices $\boldsymbol{\Pi}_j = -\boldsymbol{\Theta}^j$ in the infinite VAR form are of the same

magnitudes as the $\boldsymbol{\Psi}_j$ coefficient matrices in the previous AR(1) example. The covariance matrices of the MA(1) at lags 0 and 1 are

$$\boldsymbol{\Gamma}(0) = \boldsymbol{\Sigma} + \boldsymbol{\Theta}\boldsymbol{\Sigma}\boldsymbol{\Theta}' = \begin{bmatrix} 8.66 & 0.76 \\ 0.76 & 2.88 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Gamma}(1) = -\boldsymbol{\Sigma}\boldsymbol{\Theta}' = \begin{bmatrix} -3.9 & 1.0 \\ -2.2 & -0.8 \end{bmatrix}$$

with corresponding correlation matrices

$$\boldsymbol{\rho}(0) = \mathbf{V}^{-1/2}\boldsymbol{\Gamma}(0)\mathbf{V}^{-1/2} = \begin{bmatrix} 1.000 & 0.152 \\ 0.152 & 1.000 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\rho}(1) = \begin{bmatrix} -0.450 & 0.200 \\ -0.441 & -0.278 \end{bmatrix}$$

The above calculations are conveniently performed in R as follows:

```
> library(MTS)
> theta1=matrix(c(0.8,-0.4,0.7,0.6),2,2)
> sig=matrix(c(4,1,1,2),2,2)
> eigen(theta1)
> PIwgt(Theta=theta1)
> m1=VARMAcov(Theta=theta1, Sigma=sig, lag=1)
> names(m1)
 [1] "autocov" "ccm"
> autocov=t(m1$autocov)
> autocorr=t(m1$ccm)
```

For the bivariate MA(1) model, it follows from the autocovariance structure that each series has a univariate MA(1) model representation as $z_{it} = (1 - \eta_i B)\varepsilon_{it}$, $\sigma^2_{\varepsilon_i} = \text{var}(\varepsilon_{it})$. From Appendix A4.3, the parameter values η_i and $\sigma^2_{\varepsilon_i}$ of the component series can be determined directly by solving the relations $\rho_{ii}(1) = -\eta_i/(1 + \eta_i^2)$, $\gamma_{ii}(0) = \sigma^2_{\varepsilon_i}(1 + \eta_i^2)$, $i = 1, 2$, which lead to the values $\eta_1 = 0.628$, $\sigma^2_{\varepsilon 1} = 6.211$, and $\eta_2 = 0.303$, $\sigma^2_{\varepsilon 2} = 2.637$, respectively.

14.3.4 Model Building for Vector MA Models

The model building tools discussed for VAR models in Section 14.2 extend in a straightforward way to moving average models. As noted, the estimated cross-covariance and cross-correlation matrices are particularly useful for specifying the model order q since from (14.3.3) the corresponding theoretical quantities are zero for lags greater than q. The partial autoregression matrices, on the other hand, would show a decaying pattern for a moving average process. The parameter estimates can be obtained using the least-squares method that is equivalent to conditional likelihood method under the normality assumption. However, analogous to the univariate case, the unknown presample values can have a larger impact on the parameter estimates for VMA models. In particular, if the true parameter values are close to the boundary of the invertibility region, the conditional likelihood approach can result in biased estimates, especially for relatively short series. Because of this, the use of the *unconditional likelihood function* is typically recommended for models with moving average terms. We return to the parameter estimation in Section 14.4.5 where the exact likelihood function is discussed for the general VARMA case.

14.4 VECTOR AUTOREGRESSIVE–MOVING AVERAGE MODELS

We now assume that the matrix $\boldsymbol{\Psi}(B)$ can be represented as the product $\boldsymbol{\Psi}(B) = \boldsymbol{\Phi}^{-1}(B)\boldsymbol{\Theta}(B)$, where $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Theta}(B)$ are the autoregressive and moving average matrix polynomials defined above. This leads to the vector model

$$(\boldsymbol{Z}_t - \boldsymbol{\mu}) - \sum_{j=1}^{p} \boldsymbol{\Phi}_j(\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}) = \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j} \tag{14.4.1}$$

where $\boldsymbol{a}_t$ again is a vector white noise process with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = E[\boldsymbol{a}_t \boldsymbol{a}_t']$. The resulting process $\{\boldsymbol{Z}_t\}$ is referred to as a vector *autoregressive–moving average*, or VARMA(p, q), process regardless of whether $\{\boldsymbol{Z}_t\}$ is stationary or not.

As for the VAR(p) model, the VARMA(p, q) process can be expressed in *structural* form by premultiplying both sides of (14.4.1) by a lower triangular matrix $\boldsymbol{\Phi}_0^{\#}$ with ones on the diagonal such that $\boldsymbol{\Phi}_0^{\#}\boldsymbol{\Sigma}\boldsymbol{\Phi}_0^{\#\prime} = \boldsymbol{\Sigma}^{\#}$ is a diagonal matrix with positive diagonal elements. This gives the following representation:

$$\boldsymbol{\Phi}_0^{\#}(\boldsymbol{Z}_t - \boldsymbol{\mu}) - \sum_{j=1}^{p} \boldsymbol{\Phi}_j^{\#}(\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}) = \boldsymbol{b}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j^{\#} \boldsymbol{b}_{t-j} \tag{14.4.2}$$

where $\boldsymbol{\Phi}_j^{\#} = \boldsymbol{\Phi}_0^{\#}\boldsymbol{\Phi}_j$, $\boldsymbol{\Theta}_j^{\#} = \boldsymbol{\Phi}_0^{\#}\boldsymbol{\Theta}_j\boldsymbol{\Phi}_0^{\#-1}$, and $\boldsymbol{b}_t = \boldsymbol{\Phi}_0^{\#}\boldsymbol{a}_t$. This model displays the concurrent dependence among the components of $\boldsymbol{Z}_t$ through the lower triangular matrix $\boldsymbol{\Phi}_0^{\#}$, with diagonal elements for $\boldsymbol{\Sigma}^{\#}$, whereas the standard or *reduced form* (14.4.1) places the concurrent relationships in the covariance matrix $\boldsymbol{\Sigma}$ of the errors. More generally, premultiplication of (14.4.1) by an arbitrary nonsingular matrix $\boldsymbol{\Phi}_0^{\#}$ yields a form similar to (14.4.2) that is useful in some cases. For example, representation of a VARMA model in this general form, but with a special structure imposed on the parameter matrices, will sometimes be more useful for model specification than the standard form (14.4.1). This is discussed further in Section 14.7.

14.4.1 Stationarity and Invertibility Conditions

The stationarity conditions for a VARMA(p, q) process are the same as for the VAR(p) process discussed in Section 14.2. Hence it can be shown that the process is stationary and has an infinite moving average representation $\boldsymbol{Z}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t-j}$ if all the roots of $\det\{\boldsymbol{\Phi}(B)\} = 0$ are greater than one in absolute value. The coefficient matrices $\boldsymbol{\Psi}_j$ are determined from the relation $\boldsymbol{\Phi}(B)\boldsymbol{\Psi}(B) = \boldsymbol{\Theta}(B)$, and satisfy the recursion

$$\boldsymbol{\Psi}_j = \boldsymbol{\Phi}_1\boldsymbol{\Psi}_{j-1} + \boldsymbol{\Phi}_2\boldsymbol{\Psi}_{j-2} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{\Psi}_{j-p} - \boldsymbol{\Theta}_j \qquad j = 1, 2, \ldots \tag{14.4.3}$$

where $\boldsymbol{\Psi}_0 = \mathbf{I}, \boldsymbol{\Psi}_j = \mathbf{0}$ for $j < 0$, and $\boldsymbol{\Theta}_j = \mathbf{0}$ for $j > q$.

Conversely, the VARMA(p, q) process is invertible with an infinite AR representation similar to (14.3.2) if all the roots of $\det\{\boldsymbol{\Theta}(B)\} = 0$ are greater than one in absolute value. The coefficient weights $\boldsymbol{\Pi}_j$ in the infinite AR representation are given by the relation

$\boldsymbol{\Theta}(B)\boldsymbol{\Pi}(B) = \boldsymbol{\Phi}(B)$, and satisfy the recursion

$$\boldsymbol{\Pi}_j = \boldsymbol{\Theta}_1\boldsymbol{\Pi}_{j-1} + \boldsymbol{\Theta}_2\boldsymbol{\Pi}_{j-2} + \cdots + \boldsymbol{\Theta}_q\boldsymbol{\Pi}_{j-q} + \boldsymbol{\Phi}_j \qquad j = 1, 2, \ldots \tag{14.4.4}$$

where $\boldsymbol{\Pi}_0 = -\mathbf{I}, \boldsymbol{\Pi}_j = \mathbf{0}$ for $j < 0$, and $\boldsymbol{\Phi}_j = \mathbf{0}$ for $j > p$.

In addition, using the moving average representation, the covariance matrices for $\boldsymbol{Z}_t$ can be written as $\boldsymbol{\Gamma}(l) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j\boldsymbol{\Sigma}\boldsymbol{\Psi}'_{j+l}, l \geq 0$. From this it follows that the covariance matrix-generating function is given by $\mathbf{G}(z) = \sum_{l=-\infty}^{\infty} \boldsymbol{\Gamma}(l)z^l = \boldsymbol{\Psi}(z^{-1})\boldsymbol{\Sigma}\boldsymbol{\Psi}(z)'$; hence, the spectral density matrix of the VARMA(p, q) process is given as in (A14.1.7) with $\boldsymbol{\Psi}(z) = \boldsymbol{\Phi}^{-1}(z)\boldsymbol{\Theta}(z)$.

14.4.2 Covariance Matrix Properties of VARMA Models

For the general stationary VARMA(p, q) process $\{\boldsymbol{Z}_t\}$, it follows from the infinite MA representation $\boldsymbol{Z}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j\boldsymbol{a}_{t-j}$ that

$$E[\boldsymbol{Z}_{t-l}\boldsymbol{a}'_{t-j}] = \begin{cases} \mathbf{0} & \text{for } j < l \\ \boldsymbol{\Psi}_{j-l}\boldsymbol{\Sigma} & \text{for } j \geq l \end{cases}$$

Therefore, it is easy to determine from (14.4.1) that the covariance matrices $\boldsymbol{\Gamma}(l) = E[(\boldsymbol{Z}_{t-l} - \boldsymbol{\mu})(\boldsymbol{Z}_t - \boldsymbol{\mu})']$ of $\{\boldsymbol{Z}_t\}$ satisfy the relations

$$\boldsymbol{\Gamma}(l) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(l-j)\boldsymbol{\Phi}'_j - \sum_{j=1}^{q} \boldsymbol{\Psi}_{j-l}\boldsymbol{\Sigma}\boldsymbol{\Theta}'_j \qquad l = 0, 1, \ldots, q \tag{14.4.5}$$

and $\boldsymbol{\Gamma}(l) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(l-j)\boldsymbol{\Phi}'_j$ for $l > q$, with the convention that $\boldsymbol{\Theta}_0 = -\mathbf{I}$. Thus, the $\boldsymbol{\Gamma}(l)$ can be evaluated in terms of the AR and MA parameter matrices $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Theta}_j$, and $\boldsymbol{\Sigma}$, using these recursions.

14.4.3 Nonuniqueness and Parameter Identifiability for VARMA Models

Although the VARMA(p, q) model appears to be a straightforward extension of the univariate ARMA(p, q) model, a number of issues are associated with this extension. For example, since each AR or MA term contributes $k \times k$ parameters, the total number of parameters in the model increases rapidly as the order increases. The overflow of parameters, whose estimates can be highly correlated, makes the interpretation of the modeled results very difficult. An additional problem that arises in the VARMA case relates to the nonuniqueness of the parameters and the lack of an identifiable model representation. This issue does not arise for the pure VAR(p) model or the pure VMA(q) model discussed earlier in this chapter. But in the vector case it is possible to have two ARMA representations, $\boldsymbol{\Phi}(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}(B)\boldsymbol{a}_t$ and $\boldsymbol{\Phi}_*(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}_*(B)\boldsymbol{a}_t$ with different parameters, that give rise to the same coefficients $\boldsymbol{\Psi}_j$ in the infinite MA representation, such that

$$\boldsymbol{\Psi}(B) = \boldsymbol{\Phi}^{-1}(B)\boldsymbol{\Theta}(B) = \boldsymbol{\Phi}_*^{-1}(B)\boldsymbol{\Theta}_*(B)$$

Thus, the two models also give rise to the same covariance matrix structure $\{\boldsymbol{\Gamma}(l)\}$ and hence the same process.

Two VARMA models with this property are said to be *observationally equivalent*, or the models are said to be *exchangeable*. As a basic example, the bivariate VARMA(1, 1)

model $(\mathbf{I} - \boldsymbol{\Phi}_* B)\boldsymbol{Z}_t = (\mathbf{I} - \boldsymbol{\Theta}_* B)\boldsymbol{a}_t$ with parameters

$$\boldsymbol{\Phi}_* = \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} \qquad \boldsymbol{\Theta}_* = \begin{bmatrix} 0 & \beta \\ 0 & 0 \end{bmatrix}$$

is observationally equivalent to both a VAR(1) model $(\mathbf{I} - \boldsymbol{\Phi}B)\boldsymbol{Z}_t = \boldsymbol{a}_t$ and a VMA(1) model $\boldsymbol{Z}_t = (\mathbf{I} - \boldsymbol{\Theta}B)\boldsymbol{a}_t$, with

$$\boldsymbol{\Phi} \equiv -\boldsymbol{\Theta} = \begin{bmatrix} 0 & (\alpha - \beta) \\ 0 & 0 \end{bmatrix}$$

since, for example, $(\mathbf{I} - \boldsymbol{\Phi}_* B)^{-1}(\mathbf{I} - \boldsymbol{\Theta}_* B) = (\mathbf{I} + \boldsymbol{\Phi}_* B)(\mathbf{I} - \boldsymbol{\Phi}_* B) = (\mathbf{I} - \boldsymbol{\Theta}\boldsymbol{B})$. Hence, the parameters $\boldsymbol{\Phi}_*$ and $\boldsymbol{\Theta}_*$ in the ARMA(1, 1) model representation are not identifiable, since the properties of the process depend only on the value of $\alpha - \beta$.

In general, observationally equivalent ARMA(p, q) representations can exist because matrix AR and MA operators could be related by a common left matrix factor $\mathbf{U}(B)$ as

$$\boldsymbol{\Phi}_*(B) = \mathbf{U}(B)\boldsymbol{\Phi}(B) \quad \text{and} \quad \boldsymbol{\Theta}_*(B) = \mathbf{U}(B)\boldsymbol{\Theta}(B)$$

but such that the orders of $\boldsymbol{\Phi}_*(B)$ and $\boldsymbol{\Theta}_*(B)$ are not increased over those of $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Theta}(B)$. This common left factor $\mathbf{U}(B)$ would cancel when $\boldsymbol{\Phi}_*^{-1}(B)\boldsymbol{\Theta}_*(B)$ is formed, resulting in the same parameter matrices in $\boldsymbol{\Psi}(\boldsymbol{B})$. A particular ARMA model specification and its parameters are said to be *identifiable* if the $\boldsymbol{\Phi}_j$ and the $\boldsymbol{\Theta}_j$ are uniquely determined by the set of impulse response matrices $\boldsymbol{\Psi}_j$ in the infinite MA representation, or equivalently by the set of covariance matrices $\boldsymbol{\Gamma}(l)$ in the stationary case.

For the mixed VARMA(p, q) model, certain conditions are needed on the matrix operators $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Theta}(B)$ to ensure uniqueness of the parameters in the ARMA representation. In addition to the stationarity and invertibility conditions, the following two conditions are sufficient for identifiability:

1. The matrices $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Theta}(B)$ have no common left factors other than unimodular ones. That is, if $\boldsymbol{\Phi}(B) = \mathbf{U}(B)\boldsymbol{\Phi}_1(B)$ and $\boldsymbol{\Theta}(B) = \mathbf{U}(B)\boldsymbol{\Theta}_1(B)$, then the common factor $\mathbf{U}(B)$ must be unimodular, that is, $\det\{\mathbf{U}(B)\}$ is a nonzero constant. When this property holds, $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Theta}(B)$ are called left-coprime.
2. With q as small as possible and p as small as possible for that q, the joint matrix $[\boldsymbol{\Phi}_p, \boldsymbol{\Theta}_q]$ must be of rank k, the dimension of $\boldsymbol{Z}_t$.

Notice that through the relation $\mathbf{U}(B)^{-1} = [1/\det\{\mathbf{U}(B)\}]\text{adj}\{\mathbf{U}(B)\}$, the operator $\mathbf{U}(B)$ is a unimodular matrix if and only if $\mathbf{U}(B)^{-1}$ is a matrix polynomial of finite order. The operator $\mathbf{U}(B) = \mathbf{I} - \boldsymbol{\Phi}_* B$ in the simple ARMA(1, 1) example above is an illustration of a unimodular matrix. For further discussion of the identifiability conditions for the VARMA(p, q) model, see, for example, Hannan and Deistler (1988, Chapter 2) and Reinsel (1997, Chapter 2).

14.4.4 Model Specification for VARMA Processes

The model specification tools discussed for VAR(p) models in Section 14.2 extend in principle to the VARMA case. This includes the examination of the cross-correlation and partial autoregression matrices as discussed by Tiao and Box (1981). Additional tools

include the information criteria for model specification examined earlier, and the use of extended cross-correlation matrices for VARMA models discussed by Tiao and Tsay (1983). However, because of the identifiability issue and the overflow of parameters in the vector case, additional model specification tools focusing on the parameter structure of the VARMA representation are now needed.

Kronecker Indices. Beyond the specification of overall orders p and q, the structure of the VARMA(p, q) model can be characterized by a set of Kronecker indices $K_1, \ldots, K_k$ and the McMillan degree $M = \sum_{i=1}^{k} K_i$ of the process. The Kronecker indices, also known as structural indices, represent the maximal row degrees of the individual equations of the VARMA model. The use of these indices leads to the specification of a VARMA process of order $p = q = \max\{K_i\}$ with certain simplifying structure in the parameter matrices $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Theta}_j$. A Kronecker index equal to K_i, in particular, implies that a VARMA representation can be constructed for the process such that the ith rows of the matrices $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Theta}_j$ are zero for $j > K_i$ and with zero constraints imposed on certain other elements of $\boldsymbol{\Phi}_j$. The resulting model is referred to as the *echelon canonical form* of the VARMA model. The set of Kronecker indices is unique for a given VARMA process and the identifiability issue discussed above is thus avoided. The echelon form structure and identifiability conditions in terms of the echelon form have been examined extensively by Hannan and Deistler (1988) and others.

The Kronecker indices can be estimated using canonical correlation analysis methods introduced by Akaike (1976) and further elaborated upon by Cooper and Wood (1982) and Tsay (1989a). These methods, which are extensions of the canonical correlation analysis procedures discussed for the univariate case in Section 6.2.4, are employed to determine the nonzero canonical correlations between the past and present values of the process, $\{\boldsymbol{Z}_{t-j}, j \geq 0\}$, and the future values $\{\boldsymbol{Z}_{t+j}, j > 0\}$. In this way, the Kronecker indices K_i can be deduced, which then provide the overall model order as well as the maximum order of the AR and MA polynomials for each individual component. Further details of this approach will be given in Section 14.7. More extensive accounts of the Kronecker index approach to model specification have been provided by Solo (1986), Reinsel (1997), Lütkepohl (2006), and Tsay (1989b, 1991, 2014), among others.

Scalar Component Models. Tiao and Tsay (1989) proposed an alternative way to identify the order structure of the VARMA model based on the concept of *scalar component models* (SCMs). This approach examines linear combinations of the observed series with the goal of arriving at a parsimonious model representation that overcomes the identification issue and that may reveal meaningful structures in the data. Using this approach, k independent linear combinations $y_{it} = \boldsymbol{v}_i' \boldsymbol{Z}_t$ of orders (p_i, q_i), $i = 1, \ldots, k$, are sought such that the orders $p_i + q_i$ are as small as possible. Given a k-dimensional VARMA(p, q) process, a nonzero linear combination $y_t = \boldsymbol{v}' \boldsymbol{Z}_t$ follows SCM(p_1, q_1) if

$$y_t - \sum_{j=1}^{p_1} \boldsymbol{v}' \boldsymbol{\Phi}_j \boldsymbol{Z}_{t-j} = \boldsymbol{v}' \boldsymbol{a}_t - \sum_{j=1}^{q_1} \boldsymbol{v}' \boldsymbol{\Theta}_j a_{t-j}$$

where $0 \leq p_1 \leq p$, $0 \leq q_1 \leq q$, and $u_t = y_t - \sum_{j=1}^{p_1} \boldsymbol{v}' \boldsymbol{\Phi}_j \boldsymbol{Z}_{t-j}$ is uncorrelated with $\boldsymbol{a}_{t-j}$ for $j > q_1$. Notice that the scalar component y_t depends only on lags 1 to p_1 of all variables

$\boldsymbol{Z}_t$, and lags 1 to q_1 of all the innovations $\boldsymbol{a}_t$. Starting from SCM(0, 0), the SCM method uses a sequence of canonical correlation tests to discover k such linear combinations.

Once such a set has been found, the specification of the ARMA structure for $\boldsymbol{Z}_t$ can be determined through the relations

$$\boldsymbol{T}\boldsymbol{Z}_t - \sum_{j=1}^{p} \boldsymbol{G}_j \boldsymbol{Z}_{t-j} = \boldsymbol{T}\boldsymbol{a}_t + \sum_{j=1}^{q} \boldsymbol{H}_j \boldsymbol{a}_{t-j} \tag{14.4.6}$$

where $\boldsymbol{T} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k]'$ is a $k \times k$ nonsingular matrix, $\boldsymbol{G}_j = \boldsymbol{T}\boldsymbol{\Phi}_j, j = 1, \ldots, p$, $\boldsymbol{H}_j = \boldsymbol{T}\boldsymbol{\Theta}_j, j = 1, \ldots, q$, $p = \max\{p_i\}$ and $q = \max\{q_i\}$. Moreover, the ith row of $\boldsymbol{G}_j$ is specified to be zero for $j > p_i$ and the ith row of $\boldsymbol{H}_j$ is zero for $j > q_i$. Premultiplication of (14.4.6) by $\boldsymbol{T}^{-1}$ thus leads to a VARMA(p, q) model for $\boldsymbol{Z}_t$ in standard form but such that the coefficient matrices $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Theta}_j$ have a reduced-rank structure. On the other hand, inserting the factor $\boldsymbol{T}^{-1}\boldsymbol{T}$ in front of the $\boldsymbol{Z}_{t-j}$ and $\boldsymbol{a}_{t-j}$ in (14.4.6) yields a VARMA(p, q) representation for the *transformed process* $\boldsymbol{Y}_t = \boldsymbol{T}\boldsymbol{Z}_t$ as

$$\boldsymbol{Y}_t - \sum_{j=1}^{p} \boldsymbol{\Phi}_j^* Y_{t-j} = \boldsymbol{e}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j^* e_{t-j}$$

where $\boldsymbol{\Phi}_j^* = \boldsymbol{G}_j\, \boldsymbol{T}^{-1} = \boldsymbol{T}\boldsymbol{\Phi}_j\, \boldsymbol{T}^{-1}, \boldsymbol{\Theta}_j^* = \boldsymbol{H}_j\, \boldsymbol{T}^{-1} = \boldsymbol{T}\boldsymbol{\Theta}_j\, \boldsymbol{T}^{-1}$, and $\boldsymbol{e}_t = \boldsymbol{T}\boldsymbol{a}_t$. This VARMA representation for the transformed process is parsimonious in the sense that the ith row of $\boldsymbol{\Phi}_j^*$ is zero for $j > p_i$ and the ith row of $\boldsymbol{\Theta}_j^*$ is zero for $j > q_i$. In addition, some elements of the ith row of $\boldsymbol{\Theta}_j^*$, for $i = 1, \ldots, q_i$, are specified to be zero to remove possible redundancy of the parameters in the AR and MA matrices. The method used to identify and eliminate redundant parameters is referred to as *the rule of elimination*.

The approach of Tiao and Tsay (1989) thus identifies the scalar component processes $\boldsymbol{Y}_t = \boldsymbol{T}\boldsymbol{Z}_t$ and their associated orders (p_j, q_j) through canonical correlation methods, and then estimates a VARMA process for the transformed variables $\boldsymbol{Y}_t$ with zero constraints imposed on some of the parameters. By comparison, the Kronecker index approach estimates Kronecker indices that lead to the echelon model form for the original series $\boldsymbol{Z}_t$ directly. Also, the scalar component allows the orders of the AR and MA polynomials to differ while the orders are the same for the Kronecker index approach. The scalar component approach may in this regard be viewed as a refinement over the Kronecker index approach.

More detailed comparisons of the Kronecker index and the SCM model specification methods are provided by Reinsel (1997) and Tsay (1989b, 1991, 2014). A comparison of the forecasting performance of models specified by the two approaches was reported by Athanasopoulos et al. (2012), who found the results for SCM more favorable. Software modeling tools are available for both methods in the MTS package in R; for details and demonstrations, see Tsay (2014).

Order Determination Using Linear Least Squares. Before we proceed to discuss parameter estimation in the next section, we will mention another method that has been considered for VARMA model specification. This is a multivariate extension of the two-stage linear least-squares regression approach presented for the univariate case by Hannan and Rissanen (1982) and briefly discussed in Section 6.2.4. At the first stage of this procedure, the VARMA model is approximated by a high-order pure VAR model and the least squares method is used to obtain an estimate $\hat{\boldsymbol{a}}_t$ of the white noise error process $\boldsymbol{a}_t$. In the second

stage, one regresses $\boldsymbol{Z}_t$ on the lagged $\boldsymbol{Z}_{t-j}$ and lagged $\hat{\boldsymbol{a}}_{t-j}$ for various combinations of p and q. A model selection criterion such as BIC is then employed to help select appropriate orders for the VARMA model. Use of this procedure may lead to one or two models that seem highly promising, which are later estimated by more efficient procedures such as the maximum likelihood method. Similar linear estimation methods have been proposed by Hannan and Kavalieris (1984), Poskitt (1992), and Lütkepohl and Poskitt (1996), among others, for determining the Kronecker index structure of the VARMA model.

14.4.5 Estimation and Model Checking for VARMA Models

Once a well-defined VARMA model has been specified, the estimation of the parameters is typically performed using maximum likelihood methods assuming normality. In the past, conditional likelihood approaches were often employed for computational convenience. In the VARMA(p, q) model, this corresponds to treating the unknown presample values of $\boldsymbol{Z}_t$ and $\boldsymbol{a}_t$ as fixed constants with the $\boldsymbol{a}_t, t = 0, \ldots, 1-q$, typically set equal to zero. However, for many mixed models with an MA operator $\boldsymbol{\Theta}(B)$ having roots near the unit circle, the conditional likelihood approach has been shown to produce estimates with poorer finite sample properties than the unconditional, or exact, ML estimates.

Various approaches to the construction of the exact Gaussian likelihood function have been considered in the literature. Earlier classical approaches to evaluate the exact likelihood were presented by Hillmer and Tiao (1979) and Nicholls and Hall (1979). Given N observations $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$, the exact likelihood of a stationary VARMA(p, q) model $\boldsymbol{\Phi}(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}(B)\boldsymbol{a}_t$ has the form

$$L = |\boldsymbol{\Sigma}|^{-N/2}|\boldsymbol{\Omega}|^{-1/2}|\mathbf{D}|^{-1/2} \exp\left\{ -\left(\frac{1}{2}\right)\left[\sum_{t=1}^{N} \hat{\boldsymbol{a}}_t'\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{a}}_t + \hat{\mathbf{a}}_*'\boldsymbol{\Omega}^{-1}\hat{\mathbf{a}}_*\right]\right\} \tag{14.4.7}$$

where $\mathbf{a}_* = (\boldsymbol{Z}_{1-p}', \ldots, \boldsymbol{Z}_0', \boldsymbol{a}_{1-q}', \ldots, \boldsymbol{a}_0')'$ denotes the vector of presample values, $\hat{\mathbf{a}}_* = E[\mathbf{a}_* | \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N]$ represents the conditional expectation of $\mathbf{a}_*$ given the data, $\boldsymbol{\Omega} = \text{cov}[\mathbf{a}_*]$ denotes the covariance matrix of $\mathbf{a}_*$, and $\mathbf{D}^{-1} = \text{cov}[\mathbf{a}_* - \hat{\mathbf{a}}_*]$. The $\hat{\mathbf{a}}_t$ satisfy the recursion

$$\hat{\boldsymbol{a}}_t = \boldsymbol{Z}_t - \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{Z}_{t-j} + \sum_{j=1}^{q} \boldsymbol{\Theta}_j \hat{\boldsymbol{a}}_{t-j} \quad t = 1, \ldots, N \tag{14.4.8}$$

where the presample values are the estimated values $\hat{\boldsymbol{Z}}_t, t = 1-p, \ldots, 0$, and $\hat{\boldsymbol{a}}_t, t = 1-q, \ldots, 0$. Details of the calculations are given in the papers referenced above. Explicit expressions for the quantities $\boldsymbol{\Omega}$, $\boldsymbol{D}$, and $\hat{\mathbf{a}}_*$ are also provided by Reinsel (1997, Section 5.3.1).

Other approaches to likelihood evaluation emphasize the innovations form of the exact likelihood and the use of the state-space model representation of the VARMA model and the associated Kalman filtering methods; see, for example, Ansley and Kohn (1983), Solo (1984a), and Shea (1987). The innovations form of the exact likelihood is

$$L = \left(\prod_{t=1}^{N} |\boldsymbol{\Sigma}_{t|t-1}|^{-1/2}\right) \exp\left\{ -\left(\frac{1}{2}\right) \sum_{t=1}^{N} \boldsymbol{a}_{t|t-1}'\boldsymbol{\Sigma}_{t|t-1}^{-1}\boldsymbol{a}_{t|t-1}\right\} \tag{14.4.9}$$

where $\boldsymbol{a}_{t|t-1} = \boldsymbol{Z}_t - \hat{\boldsymbol{Z}}_{t|t-1}$ is the one-step prediction error, or innovation,

$$\hat{\boldsymbol{Z}}_{t|t-1} = E[\boldsymbol{Z}_t | \boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_1]$$

denotes the linear predictor of $\boldsymbol{Z}_t$ based on $\boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_1$, and $\boldsymbol{\Sigma}_{t|t-1} = \text{cov}[\boldsymbol{a}_{t|t-1}]$ is the one-step prediction error covariance matrix. The $\boldsymbol{a}_{t|t-1}$ and $\boldsymbol{\Sigma}_{t|t-1}$, for $t = 1, \ldots, N$, can be computed recursively using the innovations algorithm described by Brockwell and Davis (1991) and Reinsel (1997). Equivalently, the quantities $\boldsymbol{a}_{t|t-1} = \boldsymbol{Z}_t - \hat{\boldsymbol{Z}}_{t|t-1}$ and $\boldsymbol{\Sigma}_{t|t-1}$ are also obtained naturally as outputs from the Kalman filtering algorithm applied to the state-space representation of the VARMA model, which is discussed in more detail in Section 14.6. Asymptotic theory of the resulting maximum likelihood estimators for VARMA models has been studied by Dunsmuir and Hannan (1976), Deistler et al. (1978), and Hannan and Deistler (1988).

Diagnostic Checking. The checking of the fitted model can be performed using the tools described for VAR models in Section 14.2.6. These include plots of the residuals against time and/or against other variables and detailed examination of the autocorrelation and cross-correlation functions of the residuals. These tools can provide valuable information about possible lack of fit and suggest directions for model improvement. Useful supplementary tools include the portmanteau test and similar statistical tests. These tools also extend to fitted models with constraints imposed on the parameter coefficient matrices (i.e., structured parameterizations), such as echelon canonical form and reduced-rank models discussed in more detail in Section 14.7. For example, the statistic Q_s will then have $k^2 s - b$ degrees of freedom in its limiting chi-squared distribution, where b denotes the number of unconstrained parameters involved in the estimation of the ARMA model coefficients $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Theta}_j$.

14.4.6 Relation of VARMA Models to Transfer Function and ARMAX Models

The relationship between a bivariate VAR(1) model and a transfer function model was mentioned in Section 14.2.1. We will now briefly examine the relationship between subcomponents in a more general VARMA(p, q) process. We begin by partitioning the k-dimensional vector process $\boldsymbol{Z}_t$ into two groups of subcomponents of dimensions k_1 and k_2, respectively, as $\boldsymbol{Z}_t = (\boldsymbol{Z}'_{1t}, \boldsymbol{Z}'_{2t})'$. The innovations vector $\boldsymbol{a}_t$ and the AR and MA matrix polynomials are partitioned accordingly as $\boldsymbol{a}_t = (\boldsymbol{a}'_{1t}, \boldsymbol{a}'_{2t})'$ and

$$\boldsymbol{\Phi}(B) = \begin{bmatrix} \boldsymbol{\Phi}_{11}(B) & \boldsymbol{\Phi}_{12}(B) \\ \boldsymbol{\Phi}_{21}(B) & \boldsymbol{\Phi}_{22}(B) \end{bmatrix} \qquad \boldsymbol{\Theta}(B) = \begin{bmatrix} \boldsymbol{\Theta}_{11}(B) & \boldsymbol{\Theta}_{12}(B) \\ \boldsymbol{\Theta}_{21}(B) & \boldsymbol{\Theta}_{22}(B) \end{bmatrix}$$

Suppose now that $\boldsymbol{\Phi}_{12}(B)$ and $\boldsymbol{\Theta}_{12}(B)$ are both identically zero, and for convenience also assume that $\boldsymbol{\Theta}_{21}(B) = 0$. The equations for the VARMA model can then be expressed in two distinct groups as

$$\boldsymbol{\Phi}_{11}(B)\boldsymbol{Z}_{1t} = \boldsymbol{\Theta}_{11}(B)\boldsymbol{a}_{1t} \tag{14.4.10a}$$

and

$$\boldsymbol{\Phi}_{22}(B)\boldsymbol{Z}_{2t} = -\boldsymbol{\Phi}_{21}(B)\boldsymbol{Z}_{1t} + \boldsymbol{\Theta}_{22}(B)\boldsymbol{a}_{2t} \tag{14.4.10b}$$

We see from these expressions that future values of the process $\boldsymbol{Z}_{1t}$ are only influenced by its own past and not by the past of $\boldsymbol{Z}_{2t}$, whereas future values of $\boldsymbol{Z}_{2t}$ are influenced by the past of both $\boldsymbol{Z}_{1t}$ and $\boldsymbol{Z}_{2t}$. Notice that even if $\boldsymbol{\Theta}_{21}(B) \neq 0$, this conclusion still holds since the additional term in (14.4.10b) would then be $\boldsymbol{\Theta}_{21}(B)\boldsymbol{a}_{1t} = \boldsymbol{\Theta}_{21}(B)\boldsymbol{\Theta}_{11}^{-1}(B)\boldsymbol{\Phi}_{11}(B)\boldsymbol{Z}_{1t}$.

In the terminology of causality from econometrics, under (14.4.10a) and (14.4.10b), the variables $\boldsymbol{Z}_{1t}$ are said to cause $\boldsymbol{Z}_{2t}$, but $\boldsymbol{Z}_{2t}$ do not cause $\boldsymbol{Z}_{1t}$. The variables $\boldsymbol{Z}_{1t}$ are referred to as *exogenous variables*, and (14.4.10b) is often referred to as an ARMAX model or ARMAX system for the output variables $\boldsymbol{Z}_{2t}$ with $\boldsymbol{Z}_{1t}$ serving as input variables. The **X** in ARMAX stands for exogenous. The model (14.4.10b) can be rewritten as

$$\boldsymbol{Z}_{2t} = \boldsymbol{\Psi}_*(B)\boldsymbol{Z}_{1t} + \boldsymbol{\Psi}_{22}(B)\mathbf{a}_{2t}$$

where

$$\boldsymbol{\Psi}_*(B) = -\boldsymbol{\Phi}_{22}^{-1}(B)\boldsymbol{\Phi}_{21}(B) \quad \text{and} \quad \boldsymbol{\Psi}_{22}(B) = \boldsymbol{\Phi}_{22}^{-1}(B)\boldsymbol{\Theta}_{22}(B)$$

This equation provides a representation for the output process $\boldsymbol{Z}_{2t}$ as a causal linear filter of the input process $\boldsymbol{Z}_{1t}$ with added unobservable noise, that is,

$$\boldsymbol{Z}_{2t} = \boldsymbol{\Psi}_*(B)\boldsymbol{Z}_{1t} + \boldsymbol{N}_t \tag{14.4.11}$$

where the noise process $\boldsymbol{N}_t$ follows a VARMA model $\boldsymbol{\Phi}_{22}(B)\mathbf{N}_t = \boldsymbol{\Theta}_{22}(B)\mathbf{a}_{2t}$. Since the ARMAX model can be viewed as a special case of the VARMA model, the methods for model building are quite similar to those used for the VARMA model. These include the use of model selection criteria and least-squares estimation methods for model specification and examination of the residuals from the fitted model for model checking. For further discussion, see, for example, Hannan and Deistler (1988, Chapter 4) and Reinsel (1997, Chapter 8).

In the special case of bivariate time series, $\boldsymbol{Z}_{1t} \equiv z_{1t}$ and $\boldsymbol{Z}_{2t} \equiv z_{2t}$ are each univariate time series. Then we see from the above that when $\Phi_{12}(B) = 0$ and $\Theta_{12}(B) = 0$, the model reduces to the structure of the ''unidirectional'' instantaneous transfer function model with z_{1t} as the ''input'' process and z_{2t} as the output, assuming independence between z_{2t} and the noise term of z_{1t}. More generally, assuming independence between $\boldsymbol{Z}_{1t}$ and $\boldsymbol{N}_t$ above, (14.4.11) can be viewed as a multivariate generalization of the univariate (single-equation) transfer function model discussed in Chapters 11 and 12.

14.5 FORECASTING FOR VECTOR AUTOREGRESSIVE–MOVING AVERAGE PROCESSES

14.5.1 Calculation of Forecasts from ARMA Difference Equation

For forecasting in the VARMA(p, q) model

$$\boldsymbol{Z}_t = \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{Z}_{t-j} + \boldsymbol{\delta} + \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j} \tag{14.5.1}$$

where $\boldsymbol{\delta} = (\mathbf{I} - \boldsymbol{\Phi}_1 - \cdots - \boldsymbol{\Phi}_p)\boldsymbol{\mu}$ for stationary processes, we assume that the white noise series $\boldsymbol{a}_t$ are mutually independent random vectors. From general principles of prediction, the predictor of a future value $\boldsymbol{Z}_{t+l}, l = 1, 2, \ldots$, based on observations available at time

$t, \{\boldsymbol{Z}_s, s \le t\}$, that yields the minimum mean squared error (MSE) matrix is given by $\hat{\boldsymbol{Z}}_t(l) = E[\boldsymbol{Z}_{t+l} | \boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots]$. So from a computational view, forecasts are determined by applying conditional expectations to both sides of the VARMA(p, q) relation

$$\boldsymbol{\Phi}(B)\boldsymbol{Z}_{t+l} = \boldsymbol{\delta} + \boldsymbol{\Theta}(B)\boldsymbol{a}_{t+l}$$

using the result that $E[\boldsymbol{a}_{t+h} | \boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots] = \boldsymbol{0}, h > 0$, since $\boldsymbol{a}_{t+h}$ is independent of present and past values of the series. Thus, forecasts $\hat{\boldsymbol{Z}}_t(l)$ can be computed recursively from the VARMA model difference equation as

$$\hat{\boldsymbol{Z}}_t(l) = \sum_{j=1}^{p} \boldsymbol{\Phi}_j \hat{\boldsymbol{Z}}_t(l-j) + \boldsymbol{\delta} - \sum_{j=l}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t+l-j} \quad l = 1, 2, \ldots, q \tag{14.5.2}$$

with $\hat{\boldsymbol{Z}}_t(l) = \sum_{j=1}^{p} \boldsymbol{\Phi}_j \hat{\boldsymbol{Z}}_t(l-j) + \boldsymbol{\delta}$, for $l > q$, where $\hat{\boldsymbol{Z}}_t(l-j) = \boldsymbol{Z}_{t+l-j}$ for $l \le j$. Note that for pure VAR models with $q = 0$

$$\hat{\boldsymbol{Z}}_t(l) = \sum_{j=1}^{p} \boldsymbol{\Phi}_j \hat{\boldsymbol{Z}}_t(l-j) + \boldsymbol{\delta}, \quad \text{for all } l = 1, 2, \ldots$$

So the p initial forecast values are completely determined by the last p observations $\boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_{t-p+1}$; hence, for AR models all forecasts depend only on these last p observations in the series.

For models that involve an MA term, in practice it is necessary to generate the white noise sequence $\boldsymbol{a}_t$ recursively from the past data $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_t$, as

$$\boldsymbol{a}_s = \boldsymbol{Z}_s - \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{Z}_{s-j} - \boldsymbol{\delta} + \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{s-j} \quad s = 1, 2, \ldots, t$$

using appropriate starting values for $\boldsymbol{a}_0, \ldots, \boldsymbol{a}_{1-q}$ and $\boldsymbol{Z}_0, \ldots, \boldsymbol{Z}_{1-p}$. One way to estimate the starting values is to use the *backcasting* technique described earlier for evaluation of the exact likelihood function for ARMA models. This method yields $\hat{\boldsymbol{a}}_{1-j} = E[\boldsymbol{a}_{1-j} | \boldsymbol{Z}_t, \ldots, \boldsymbol{Z}_1]$, $j = 1, \ldots, q$, and $\hat{\boldsymbol{Z}}_{1-j} = E[\boldsymbol{Z}_{1-j} | \boldsymbol{Z}_t, \ldots, \boldsymbol{Z}_1], j = 1, \ldots, p$. The resulting forecasts $\hat{\boldsymbol{Z}}_t(l)$ are then equal to

$$\hat{\boldsymbol{Z}}_t(l) \equiv E[\boldsymbol{Z}_{t+l} | \boldsymbol{Z}_t, \ldots, \boldsymbol{Z}_1]$$

These are optimal forecasts based on the finite past history $\boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_1$, although the analysis of forecast properties given below assumes that the forecasts are based on the infinite past history $\boldsymbol{Z}_s$, all $s \le t$. However, these two forecasts will be nearly identical for any moderate or large value of t, the number of past values available for forecasting. Alternative methods to obtain the ''exact'' finite sample forecasts, as well as the exact covariance matrices of the forecast errors, based on the finite sample data $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t$, in a convenient computational manner are through an innovations approach or through the closely related state-space model and Kalman filter approach that will be discussed briefly in Section 14.6.

14.5.2 Forecasts from Infinite VMA Form and Properties of Forecast Errors

To establish the theoretical MSE properties of the forecast errors, we use the ''infinite'' moving average representation $\boldsymbol{Z}_t = \boldsymbol{\Psi}(B)\boldsymbol{a}_t$ of the VARMA(p, q) model, where $\boldsymbol{\Psi}(B) = \boldsymbol{\Phi}^{-1}(B)\boldsymbol{\Theta}(B) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j B^j$. A future value $\boldsymbol{Z}_{t+l}$, relative to the forecast origin t, can then be expressed as

$$\boldsymbol{Z}_{t+l} = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t+l-j} = \boldsymbol{a}_{t+l} + \boldsymbol{\Psi}_1 \boldsymbol{a}_{t+l-1} + \cdots + \boldsymbol{\Psi}_{l-1} \boldsymbol{a}_{t+1} + \boldsymbol{\Psi}_l \boldsymbol{a}_t + \cdots$$

Thus, since $E[\boldsymbol{a}_{t+h} | \boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots] = 0$, $h > 0$, the minimum MSE matrix predictor of $\boldsymbol{Z}_{t+l}$ based on $\boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots$ can be represented as

$$\hat{\boldsymbol{Z}}_t(l) = E[\boldsymbol{Z}_{t+l} | \boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots] = \sum_{j=l}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t+l-j} \tag{14.5.3}$$

The l-step-ahead forecast error is $\boldsymbol{e}_t(l) = \boldsymbol{Z}_{t+l} - \hat{\boldsymbol{Z}}_t(l) = \sum_{j=0}^{l-1} \boldsymbol{\Psi}_j \boldsymbol{a}_{t+l-j}$ has zero mean and covariance matrix:

$$\boldsymbol{\Sigma}(l) = \text{cov}[\boldsymbol{e}_t(l)] = E[\boldsymbol{e}_t(l)\boldsymbol{e}_t(l)'] = \sum_{j=0}^{l-1} \boldsymbol{\Psi}_j \boldsymbol{\Sigma} \boldsymbol{\Psi}_j' \quad \boldsymbol{\Psi}_0 = \mathbf{I} \tag{14.5.4}$$

In particular, for one step ahead, $\boldsymbol{e}_t(1) = \boldsymbol{Z}_{t+1} - \hat{\boldsymbol{Z}}_t(1) = \boldsymbol{a}_{t+1}$ with error covariance matrix $\boldsymbol{\Sigma}$, so that the white noise series $\boldsymbol{a}_t$ can be interpreted as a sequence of one-step-ahead forecast errors for the process.

It follows from the infinite MA representation of the forecasts given by (14.5.3) that we obtain the multivariate version of the updating formula (5.2.5) as

$$\hat{\boldsymbol{Z}}_{t+1}(l) = E[\boldsymbol{Z}_{t+l+1} | \boldsymbol{Z}_{t+1}, \boldsymbol{Z}_t, \ldots] = \sum_{j=l}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t+l+1-j} = \hat{\boldsymbol{Z}}_t(l+1) + \boldsymbol{\Psi}_l \boldsymbol{a}_{t+1} \tag{14.5.5}$$

where $\boldsymbol{a}_{t+1} = \boldsymbol{Z}_{t+1} - \hat{\boldsymbol{Z}}_t(1)$ is the one-step-ahead forecast error. This provides a simple relationship to indicate how the forecast $\hat{\boldsymbol{Z}}_t(l)$ with forecast origin t is adjusted or updated to incorporate the information available from a new observation $\boldsymbol{Z}_{t+1}$ at time $t + 1$.

For the case of unit-root nonstationary processes to be discussed in Section 14.8, similar forecasting topics as presented above can also be developed and results such as (14.5.2) and (14.5.4) continue to apply.

14.6 STATE-SPACE FORM OF THE VARMA MODEL

The state-space model was introduced for univariate ARMA models in Section 5.5. Similar to the univariate case, the VARMA model can be represented in the equivalent state-space form, which is of interest for purposes of prediction as well as for model specification and maximum likelihood estimation of parameters. The state-space model consists of a transition or state equation

$$\boldsymbol{Y}_t = \boldsymbol{\Phi} \boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t$$

and an observation equation

$$\boldsymbol{Z}_t = \mathbf{H}\boldsymbol{Y}_t + \boldsymbol{N}_t$$

where $\boldsymbol{Y}_t$ is an $r \times 1$ (unobservable) time series vector called the state vector, and $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{N}_t$ are independent white noise processes. In this representation, the state vector $\boldsymbol{Y}_t$ conceptually contains all information from the past of the process $\boldsymbol{Z}_t$, which is relevant for the future of the process, and, hence, the dynamics of the system can be represented in the simple first-order or Markovian transition equation for the state vector. The above state-space model is said to be stable if all the eigenvalues of the matrix $\boldsymbol{\Phi}$ are less than one in absolute value, and conversely, it can be shown that any stationary process $\boldsymbol{Z}_t$ that has a stable state-space representation of the above form can also be represented in the form of a stationary VARMA(p, q) model; see, for example, Akaike (1974b). Hence, it follows that any process $\boldsymbol{Z}_t$ that satisfies a stable state-space representation can be expressed in the causal convergent infinite moving average form $\boldsymbol{Z}_t = \boldsymbol{\Psi}(B)\boldsymbol{a}_t$. The stability condition for the matrix $\boldsymbol{\Phi}$ in the state-space model is equivalent to the stability condition for the matrix coefficients $\boldsymbol{\Psi}_j$ of the linear filter $\boldsymbol{\Psi}(B)$ (see Appendix A14.1.2), since it ensures that $\sum_{j=0}^{\infty} ||\boldsymbol{\Psi}_j|| < \infty$ in the representation $\boldsymbol{Z}_t = \boldsymbol{\Psi}(B)\boldsymbol{a}_t$.

For the VARMA(p, q) model (14.5.1) (with $\boldsymbol{\delta} = \mathbf{0}$), define the predictors $\hat{\boldsymbol{Z}}_t(j) = E[\boldsymbol{Z}_{t+j}|\boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots]$ as in Section 14.5.1 for $j = 0, 1, \ldots, r-1$, with $r = \max(p, q+1)$, and $\hat{\boldsymbol{Z}}_t(0) = \boldsymbol{Z}_t$. From the updating equations (14.5.5), we have $\hat{\boldsymbol{Z}}_t(j-1) = \hat{\boldsymbol{Z}}_{t-1}(j) + \boldsymbol{\Psi}_{j-1}\boldsymbol{a}_t$, $j = 1, 2, \ldots, r-1$. Also, for $j = r > q$ we find using (14.5.2) that

$$\hat{\boldsymbol{Z}}_t(j-1) = \hat{\boldsymbol{Z}}_{t-1}(j) + \boldsymbol{\Psi}_{j-1}\boldsymbol{a}_t = \sum_{i=1}^{p} \boldsymbol{\Phi}_i \hat{\boldsymbol{Z}}_{t-1}(j-i) + \boldsymbol{\Psi}_{j-1}\boldsymbol{a}_t$$

Let us define the ''state'' vector at time t, with r vector components, as $\boldsymbol{Y}_t = [\hat{\boldsymbol{Z}}_t(0)', \hat{\boldsymbol{Z}}_t(1)', \ldots, \hat{\boldsymbol{Z}}_t(r-1)']'$. Then, from the relations above, the state vector $\boldsymbol{Y}_t$ satisfies the state-space (transition) equations

$$\boldsymbol{Y}_t = \begin{bmatrix} \mathbf{0} & \mathbf{I} & \mathbf{0} \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdot \cdots & \mathbf{I} \\ \boldsymbol{\Phi}_r & \boldsymbol{\Phi}_{r-1} & \cdot \cdots & \boldsymbol{\Phi}_1 \end{bmatrix} \boldsymbol{Y}_{t-1} + \begin{bmatrix} \mathbf{I} \\ \boldsymbol{\Psi}_1 \\ \vdots \\ \boldsymbol{\Psi}_{r-2} \\ \boldsymbol{\Psi}_{r-1} \end{bmatrix} \boldsymbol{a}_t \tag{14.6.1}$$

where $\boldsymbol{\Phi}_i = \mathbf{0}$ if $i > p$. Thus, we have

$$\boldsymbol{Y}_t = \boldsymbol{\Phi}\boldsymbol{Y}_{t-1} + \boldsymbol{\Psi}\boldsymbol{a}_t \tag{14.6.2}$$

together with the observation equation

$$\boldsymbol{Z}_t^* = \boldsymbol{Z}_t + \boldsymbol{N}_t = [\mathbf{I}, \mathbf{0}, \ldots, \mathbf{0}]\boldsymbol{Y}_t + \boldsymbol{N}_t = \mathbf{H}\boldsymbol{Y}_t + \boldsymbol{N}_t \tag{14.6.3}$$

where the vector noise $\boldsymbol{N}_t$ would be present only if the process $\boldsymbol{Z}_t$ is observed subject to additional white noise error; otherwise, we simply have $\boldsymbol{Z}_t = \boldsymbol{Z}_t^* = \mathbf{H}\boldsymbol{Y}_t$. For convenience, we assume in the remainder of this section that the additional white noise is not present.

The state or transition equation (14.6.2) and the observation equation (14.6.3) constitute a state-space representation of the VARMA model. There are many other constructions

of the state vector $\boldsymbol{Y}_t$ that will give rise to state-space equations of the general form (14.6.2) and (14.6.3); that is, the state-space form of the VARMA model is not unique. Specifically, if we transform the state vector $\boldsymbol{Y}_t$ into $\overline{\boldsymbol{Y}}_t = \mathbf{P}\boldsymbol{Y}_t$, where $\mathbf{P}$ is an arbitrary nonsingular matrix, then models (14.6.2) and (14.6.3) can be written in a similar form in terms of $\overline{\boldsymbol{Y}}_t$ with $\bar{\boldsymbol{\Phi}} = \mathbf{P}\boldsymbol{\Phi}\mathbf{P}^{-1}, \overline{\mathbf{H}} = \mathbf{H}\mathbf{P}^{-1}$, and $\overline{\boldsymbol{\Psi}} = \mathbf{P}\boldsymbol{\Psi}$. The particular form given above has the state vector $\boldsymbol{Y}_t$, which can be viewed as generating the space of predictions of all future values of the process $\boldsymbol{Z}_t$, since $\hat{\boldsymbol{Z}}_t(l) = \sum_{i=1}^{r} \boldsymbol{\Phi}_i \hat{\boldsymbol{Z}}_t(l-i)$ for $l > r-1$.

In the state-space model, the unobservable state vector $\boldsymbol{Y}_t$ constitutes a summary of the state of the dynamic system through time t, and the state equation (14.6.2) describes the evolution of the dynamic system in time. The minimal dimension of the state vector $\boldsymbol{Y}_t$ in a state-space representation needs to be sufficiently large so that the dynamics of the system can be represented by the simple Markovian first-order structure. State-space representations for the VARMA model can exist with a state vector of minimal dimension smaller than the dimension in (14.6.1). This minimal dimension is the dimension of the set of basis predictors that generate the linear space of predictors of all future values; it is of smaller dimension than in (14.6.1) whenever the state vector $\boldsymbol{Y}_t$ can be represented linearly in terms of a smaller number of basis elements. Specifically, suppose that $\boldsymbol{Y}_t$ in (14.6.1) can be expressed as $\boldsymbol{Y}_t = \mathbf{A}\boldsymbol{Y}_t^*$, where $\boldsymbol{Y}_t^*$ is an $M \times 1$ vector whose elements form a subset of the elements of $\boldsymbol{Y}_t$, with $M < rk$ being the smallest possible such dimension. Then $\mathbf{A}$ is a $rk \times M$ matrix of full rank M, with $\boldsymbol{Y}_t^* = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\boldsymbol{Y}_t$, and we assume the first $k \times M$ block row of $\mathbf{A}$ is $[\mathbf{I}, \mathbf{0}, \ldots, \mathbf{0}]$. Thus, multiplying (14.6.2) on the left by $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$, we obtain the equivalent representation of minimal dimension M given by $\boldsymbol{Y}_t^* = \boldsymbol{\Phi}^*\boldsymbol{Y}_{t-1}^* + \boldsymbol{\Psi}^*\boldsymbol{a}_t$, where $\boldsymbol{\Phi}^* = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\boldsymbol{\Phi}\mathbf{A}$ and $\boldsymbol{\Psi}^* = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\boldsymbol{\Psi}$, with $\boldsymbol{Z}_t = \mathbf{H}\mathbf{A}\boldsymbol{Y}_t^* \equiv \mathbf{H}\boldsymbol{Y}_t^*$. This minimal dimension M is in fact the McMillan degree of the process $\{\boldsymbol{Z}_t\}$ as described in Section 14.7.1 below.

One important use of the state-space form of the VARMA model is that it enables exact finite sample forecasts of the process $\{\boldsymbol{Z}_t\}$ to be obtained through Kalman filtering and the associated prediction algorithm. This provides a convenient computational procedure to obtain the minimum MSE matrix estimate of the state vector $\boldsymbol{Y}_{t+l}$ based on observations $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t$ as $\hat{\boldsymbol{Y}}_{t+l|t} = E[\boldsymbol{Y}_{t+l} | \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t]$, with

$$\mathbf{P}_{t+l|t} = E[(\boldsymbol{Y}_{t+l} - \hat{\boldsymbol{Y}}_{t+l|t})(\boldsymbol{Y}_{t+l} - \hat{\boldsymbol{Y}}_{t+l|t})']$$

equal to the error covariance matrix. The recursions for the Kalman filter procedure have been presented as equations (5.5.6) to (5.5.9) in Section 5.5.2. It follows that optimal forecasts $\hat{\boldsymbol{Z}}_{t+l|t} = E[\boldsymbol{Z}_{t+l} | \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t]$ of future observations $\boldsymbol{Z}_{t+l}$ are then available as $\hat{\boldsymbol{Z}}_{t+l|t} = \mathbf{H}\hat{\boldsymbol{Y}}_{t+l|t}$, since $\boldsymbol{Z}_{t+l} = \mathbf{H}\boldsymbol{Y}_{t+l}$, with forecast error covariance matrix

$$\boldsymbol{\Sigma}_{t+l|t} = E[(\boldsymbol{Z}_{t+l} - \hat{\boldsymbol{Z}}_{t+l|t})(\boldsymbol{Z}_{t+l} - \hat{\boldsymbol{Z}}_{t+l|t})'] = \mathbf{H}\mathbf{P}_{t+l|t}\mathbf{H}'$$

The ''steady-state'' values of the Kalman filtering lead l forecast error covariance matrices, obtained as t increases, equal the expressions in (14.5.4) of Section 14.5.2, $\boldsymbol{\Sigma}(l) = \sum_{j=0}^{l-1} \boldsymbol{\Psi}_j \boldsymbol{\Sigma} \boldsymbol{\Psi}_j'$. That is, $\boldsymbol{\Sigma}_{t+l|t}$ approaches $\boldsymbol{\Sigma}(l)$ as $t \to \infty$.

Thus, the Kalman filtering procedure provides a convenient method to obtain exact finite sample forecasts for future values in the VARMA process, based on observations $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t$, subject to specification of appropriate initial conditions to use in (5.5.6) to (5.5.9). In particular, for the VARMA process represented in state-space form, the

exact finite-sample one-step-ahead forecasts $\hat{\boldsymbol{Z}}_{t|t-1} = \mathbf{H}\hat{\boldsymbol{Y}}_{t|t-1}$, and their error covariance matrices $\boldsymbol{\Sigma}_{t|t-1} = \mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}'$, can be obtained conveniently through the Kalman filtering equations. This can be particularly useful for evaluation of the exact Gaussian likelihood function, based on N vector observations $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ from the VARMA process, as mentioned earlier in Section 14.4.5.

14.7 FURTHER DISCUSSION OF VARMA MODEL SPECIFICATION

In this section, we return to the issue of model specification for a vector ARMA process. As noted in Section 14.4.4, extending the ARMA model to the vector case involves some difficulties that are not present in the univariate case. One problem in the vector case is the overflow of parameters, whose estimates can be highly correlated. A second issue is that of identifiability, which refers to the fact that two different sets of parameters can give rise to the same probability structure and hence the same process. This causes problems at the parameter estimation stage, in particular, since the likelihood function will not have a uniquely defined maximum in this case. Two methods designed to overcome these issues are the Kronecker index approach that originates in the engineering literature and the SCM method developed by Tiao and Tsay (1989). Both methods make use of canonical correlation analysis methods to arrive at a parsimonious and well-defined VARMA model.

In this section, we will discuss the VARMA model specification in more detail focusing on the Kronecker index approach to model specification. We first discuss the estimation of the Kronecker indices and the McMillan degree of a vector process. We then describe the specification of the echelon canonical form of the VARMA model through the Kronecker indices. A brief discussion of the use of partial canonical correlation analysis to identify models with reduced rank structure is also included.

14.7.1 Kronecker Structure for VARMA Models

The VARMA(p, q) model (14.4.1) can always be expressed in the equivalent form

$$\boldsymbol{\Phi}_0^{\#}(\boldsymbol{Z}_t - \boldsymbol{\mu}) - \sum_{j=1}^{p} \boldsymbol{\Phi}_j^{\#}(\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}) = \boldsymbol{\Theta}_0^{\#}\boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j^{\#}\boldsymbol{a}_{t-j} \tag{14.7.1}$$

where $\boldsymbol{\Phi}_0^{\#}$ is an arbitrary nonsingular matrix, $\boldsymbol{\Phi}_j^{\#} = \boldsymbol{\Phi}_0^{\#}\boldsymbol{\Phi}_j$, $\boldsymbol{\Theta}_0^{\#} = \boldsymbol{\Phi}_0^{\#}$, and $\boldsymbol{\Theta}_j^{\#} = \boldsymbol{\Phi}_0^{\#}\boldsymbol{\Theta}_j$. For purposes of parsimony, we are interested in model forms that lead to the simplest structure in some sense, such as in terms of the number of unknown parameters in the matrices $\boldsymbol{\Phi}_0^{\#}, \boldsymbol{\Phi}_1^{\#}, \ldots, \boldsymbol{\Phi}_p^{\#}, \boldsymbol{\Theta}_1^{\#}, \ldots, \boldsymbol{\Theta}_q^{\#}$. For unique identifiability of the parameters, it is necessary to normalize the form of $\boldsymbol{\Phi}_0^{\#}$ at least to be lower triangular with ones on the diagonal.

As discussed in detail by Hannan and Deistler (1988, Chapter 2), a representation of a VARMA model in a certain special form of (14.7.1) can sometimes be more useful for model specification than the standard or reduced VARMA form (14.4.1), and this form of (14.7.1) is referred to as the *echelon canonical form* of the VARMA model. To specify the echelon canonical form, k *Kronecker indices* or structural indices, $K_1, \ldots, K_k$, must be determined beyond the overall orders p and q. The echelon (canonical) form is such that $[\boldsymbol{\Phi}^{\#}(B), \boldsymbol{\Theta}^{\#}(B)]$ has the smallest possible row degrees, and K_i denotes the degree of the ith row of $[\boldsymbol{\Phi}^{\#}(B), \boldsymbol{\Theta}^{\#}(B)]$, that is, the maximum of the degrees of the polynomials

in the ith row of $[\boldsymbol{\Phi}^{\#}(B),\ \boldsymbol{\Theta}^{\#}(B)]$, for $i = 1, \ldots, k$, and with $p = q = \max\{K_1, \ldots, K_k\}$. The specification of these Kronecker indices or ''row orders'' $\{K_i\}$, which are unique for any given equivalence class of ARMA models, that is, models with the same infinite MA operator $\boldsymbol{\Psi}(B)$, then determines a unique echelon canonical form of the VARMA model (14.7.1) in which the unknown parameters are uniquely identifiable.

Kronecker Indices and McMillan Degree of VARMA Process. For any stationary vector process $\{\boldsymbol{Z}_t\}$ with covariance matrices $\boldsymbol{\Gamma}(l) = \text{cov}[\boldsymbol{Z}_t, \boldsymbol{Z}_{t+l}]$, we define the infinite-dimensional (block) Hankel matrix of the covariances as

$$\mathbf{H} = \begin{bmatrix} \boldsymbol{\Gamma}(1)' & \boldsymbol{\Gamma}(2)' & \boldsymbol{\Gamma}(3)' & \cdots \\ \boldsymbol{\Gamma}(2)' & \boldsymbol{\Gamma}(3)' & \boldsymbol{\Gamma}(4)' & \cdots \\ \boldsymbol{\Gamma}(3)' & \boldsymbol{\Gamma}(4)' & \boldsymbol{\Gamma}(5)' & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \tag{14.7.2}$$

Then, in particular, the *McMillan degree* M of the process is defined as the rank of the Hankel matrix $\mathbf{H}$. The process $\{\boldsymbol{Z}_t\}$ follows a finite-order VARMA model if and only if the rank of $\mathbf{H}$ is finite. For a stationary VARMA(p, q) process, the moment relations (14.4.5) yield that

$$\boldsymbol{\Gamma}(l)' - \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{\Gamma}(l-j)' = \mathbf{0} \qquad \text{for } l > q \tag{14.7.3}$$

It can be seen directly from this that the rank of $\mathbf{H}$, the McMillan degree M, will then satisfy $M \leq ks$, where $s = \max\{p, q\}$, since all the $k \times k$ block rows of $\mathbf{H}$ beyond the sth block row will be linearly dependent on the preceding block rows. But the McMillan degree M of a VARMA(p, q) could be considerably smaller than ks due to rank deficiencies in the AR and MA coefficient matrices.

The McMillan degree M has the interpretation as the number of linearly independent linear combinations of the present and past vectors $\boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots$ that are needed for optimal prediction of all future vectors within the ARMA structure. Note that

$$\mathbf{H} = \text{cov}[\mathbf{F}_{t+1}, \mathbf{P}_t] = \text{cov}[\mathbf{F}_{t+1|t}, \mathbf{P}_t] \tag{14.7.4}$$

is the covariance between the collection of all present and past vectors, $\mathbf{P}_t = (\boldsymbol{Z}_t', \boldsymbol{Z}_{t-1}', \ldots)'$, and the collection of all future vectors $\mathbf{F}_{t+1} = (\boldsymbol{Z}_{t+1}', \boldsymbol{Z}_{t+2}', \ldots)'$ or the collection of predicted values of all future vectors, $\mathbf{F}_{t+1|t} = E[\mathbf{F}_{t+1}|\mathbf{P}_t]$. Hence, if the rank of $\mathbf{H}$ is equal to M, then the (linear) predictor space formed from the collection $\mathbf{F}_{t+1|t}$ of predicted values $\hat{\boldsymbol{Z}}_t(l) = E[\boldsymbol{Z}_{t+l}|\mathbf{P}_t]$, $l > 0$, of all future vectors is of finite dimension M. Sometimes (e.g., Hannan and Deistler, 1988, Chapter 2) the Hankel matrix $\mathbf{H}$ is defined in terms of the coefficients $\boldsymbol{\Psi}_j$ in the infinite MA form $\boldsymbol{Z}_t - \boldsymbol{\mu} = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \mathbf{a}_{t-j}$ of the ARMA process, instead of the covariance matrices $\boldsymbol{\Gamma}(j)'$, but all main conclusions hold in either case.

In addition, the ith *Kronecker index* K_i, $i = 1, \ldots, k$, of the process $\{\boldsymbol{Z}_t\}$ is the smallest value such that the $(kK_i + i)$th row of $\mathbf{H}$, that is, the ith row in the $(K_i + 1)$th block of rows of $\mathbf{H}$, is linearly dependent on the previous rows of $\mathbf{H}$. This also implies, through the structure of the Hankel matrix $\mathbf{H}$, that all rows $kl + i$, for every $l \geq K_i$, will also be linearly dependent on the rows preceding the $(kK_i + i)$th row. The set of Kronecker indices $\{K_1, \ldots, K_k\}$ is unique for any given VARMA process; hence, it is not dependent

on any one particular form of the observationally equivalent ARMA model representations of the process. As indicated in Section 14.6, the VARMA model can be represented in its equivalent minimal dimension state-space form, with minimal dimension, the McMillan degree

$$M = \sum_{i=1}^{k} K_i = K_1 + K_2 + \cdots + K_k$$

being the number of linearly independent predictors required to generate the linear prediction space $\{\hat{\boldsymbol{Z}}_t(l),\ l \geq 1\}$ of all future vectors $\{\boldsymbol{Z}_{t+l}, l \geq 1\}$. This minimal dimension state-space representation is one way to reveal the special structure of the VARMA parameters associated with the Kronecker indices. Canonical correlation analysis methods between past and future vectors of a VARMA process $\{\boldsymbol{Z}_t\}$ are useful as a means to determine the Kronecker indices of the process. We will now indicate, in particular, the direct connections that the Kronecker indices have with the second moment equations as in (14.4.5) and (14.7.3), since these equations exhibit the row dependencies among the covariance matrices $\boldsymbol{\Gamma}(j)'$. Hence, knowledge of these Kronecker indices can be used to deduce special structure among the AR and MA parameter matrices and lead to specification of the special (echelon) form of the VARMA model.

Echelon Canonical Form Implied by Kronecker Indices. Specifically, if VARMA models similar to the form in (14.7.1) are considered, with $\boldsymbol{\Phi}_0^{\#} = \boldsymbol{\Theta}_0^{\#}$ lower triangular (and having ones on the diagonal), then equations similar to (14.4.5) for the cross-covariance matrices $\boldsymbol{\Gamma}(l)$ of the process are obtained as

$$\boldsymbol{\Phi}_0^{\#}\boldsymbol{\Gamma}(l)' - \sum_{j=1}^{p} \boldsymbol{\Phi}_j^{\#}\boldsymbol{\Gamma}(l-j)' = -\sum_{j=l}^{q} \boldsymbol{\Theta}_j^{\#}\boldsymbol{\Sigma}\boldsymbol{\Psi}_{j-l}' \tag{14.7.5}$$

Thus, if $\boldsymbol{\phi}_j(i)'$ denotes the ith row of $\boldsymbol{\Phi}_j^{\#}$, then the ith Kronecker index equal to K_i implies the linear dependence in the rows of the Hankel matrix $\mathbf{H}$ of the form

$$\boldsymbol{\phi}_0(i)'\boldsymbol{\Gamma}(l)' - \sum_{j=1}^{K_i} \boldsymbol{\phi}_j(i)'\boldsymbol{\Gamma}(l-j)' = \mathbf{0}' \qquad \text{for all } l \geq K_i + 1 \tag{14.7.6}$$

that is, $\boldsymbol{b}_i'\mathbf{H} = \mathbf{0}'$ with $\boldsymbol{b}_i' = (-\boldsymbol{\phi}_{K_i}(i)', \ldots, -\boldsymbol{\phi}_1(i)',\ \boldsymbol{\phi}_0(i)', \mathbf{0}', \ldots)$. Note that by definition of the ith Kronecker index K_i, the row vector $\boldsymbol{\phi}_0(i)'$ in (14.7.6) can be taken to have a one in the ith position and zeros for positions greater than the ith. Therefore, a Kronecker index equal to K_i implies, in particular, that an ARMA model representation of the form (14.7.1) can be constructed for the process such that the ith rows of the matrices $\boldsymbol{\Phi}_j^{\#}$ and $\boldsymbol{\Theta}_j^{\#}$ will be zero for $j > K_i$.

In addition to these implications from (14.7.6), additional zero constraints on certain elements in the ith rows of the matrices $\boldsymbol{\Phi}_j^{\#}$ for $j \leq K_i$ can be specified. Specifically, the lth element of the ith row $\boldsymbol{\phi}_j(i)'$ can be specified to be zero whenever $j + K_l \leq K_i$ because for $K_l \leq K_i$ the rows $k(K_l + j) + l,\ j = 0, \ldots, (K_i - K_l)$, of the Hankel matrix $\mathbf{H}$ are all

linearly dependent on the previous rows of $\mathbf{H}$. Hence, the (i, l)th element of the AR operator

$$\mathbf{\Phi}^{\#}(B) = \mathbf{\Phi}_0^{\#} - \sum_{j=1}^{p} \mathbf{\Phi}_j^{\#} B^j$$

in model (14.7.1) can be specified to have nonzero coefficients only for the lags $j = K_i - K_{il} + 1, \ldots, K_i$, with zero coefficients specified for any lower lags of j (when $i \neq l$), where we define

$$K_{il} = \begin{cases} \min(K_i + 1, K_l) & \text{for } i > l \\ \min(K_i, K_l) & \text{for } i \leq l \end{cases} \tag{14.7.7}$$

(so that whenever $K_l \leq K_i$ we have $K_{il} = K_l$). Thus, the corresponding number of unknown AR parameters in the (i, l)th element of $\mathbf{\Phi}^{\#}(B)$ is equal to K_{il}. Hence, the AR operator $\mathbf{\Phi}^{\#}(B)$ in model (14.7.1) can be specified such that the total number of unknown parameters of $\mathbf{\Phi}^{\#}(B)$ is equal to $\sum_{i=1}^{k} \sum_{l=1}^{k} K_{il} = M + \sum \sum_{i \neq l}^{k} K_{il}$, while the number of unknown parameters in the MA operator $\mathbf{\Theta}^{\#}(B)$, excluding those parameters in $\mathbf{\Theta}_0^{\#} = \mathbf{\Phi}_0^{\#}$, is equal to $\sum_{i=1}^{k} kK_i = kM$.

In summary, for a stationary linear process $\{\boldsymbol{Z}_t\}$ with Kronecker indices $K_1, \ldots, K_k$, a VARMA representation as in (14.7.1) with $p = q = \{\max K_i\}$ can be specified to describe the process, with the matrices $\mathbf{\Phi}_j^{\#}$ and $\mathbf{\Theta}_j^{\#}$ possessing the structure that their ith rows are zero for $j > K_i$ and the additional zero constraints structure noted above. Moreover, for a stationary vector process with given covariance matrix structure $\mathbf{\Gamma}(l)$, or equivalently with given infinite MA coefficients $\mathbf{\Psi}_j$, Hannan and Deistler (1988, Theorem 2.5.1) have shown that this model provides a unique VARMA representation, with AR and MA operators $\mathbf{\Phi}^{\#}(B)$ and $\mathbf{\Theta}^{\#}(B)$ being left-coprime, and where all unknown parameters are identified. This (canonical) ARMA representation is referred to as a (reversed) *echelon ARMA form*. In particular, the VAR coefficient matrices $\mathbf{\Phi}_j^{\#}$ in the echelon canonical representation (14.7.1) are uniquely determined from the $\mathbf{\Gamma}(l)$ by the requirement that their ith rows $\phi_j(i)'$, $j = 0, \ldots, K_i, i = 1, \ldots, k$, satisfy the conditions (14.7.6).

Examples. For simple illustrative examples, consider a bivariate ($k = 2$) process $\{\boldsymbol{Z}_t\}$. When this process has Kronecker indices $K_1 = K_2 = 1$, then a general VARMA(1, 1) representation $\boldsymbol{Z}_t - \mathbf{\Phi}_1 \boldsymbol{Z}_{t-1} = \boldsymbol{a}_t - \mathbf{\Theta}_1 \boldsymbol{a}_{t-1}$ is implied. However, notice that a pure VAR(1) process with full-rank VAR matrix $\mathbf{\Phi}_1$ and a pure VMA(1) process with full-rank VMA matrix $\mathbf{\Theta}_1$ would both also possess Kronecker indices equal to $K_1 = K_2 = 1$. This simple example thus illustrates that specification of the Kronecker indices alone does not necessarily lead to the specification of a VARMA representation where all the simplifying structure in the parameters is directly revealed. For a second case, suppose the bivariate process has Kronecker indices $K_1 = 1$ and $K_2 = 0$. Then, the implied structure for the process is VARMA(1, 1) as in (14.7.1), with (note, in particular, that $K_{12} = 0$ in (14.7.7))

$$\mathbf{\Phi}_0^{\#} = \begin{bmatrix} 1 & 0 \\ X & 1 \end{bmatrix} \quad \mathbf{\Phi}_1^{\#} = \begin{bmatrix} X & 0 \\ 0 & 0 \end{bmatrix} \quad \mathbf{\Theta}_1^{\#} = \begin{bmatrix} X & X \\ 0 & 0 \end{bmatrix}$$

where the X's denote unknown parameters that need estimation and 0's indicate values that are known to be specified as zero. On multiplication of the VARMA(1, 1) relation $\mathbf{\Phi}_0^{\#} \boldsymbol{Z}_t - \mathbf{\Phi}_1^{\#} \boldsymbol{Z}_{t-1} = \mathbf{\Theta}_0^{\#} \boldsymbol{a}_t - \mathbf{\Theta}_1^{\#} \boldsymbol{a}_{t-1}$ on the left by $\mathbf{\Phi}_0^{\#-1}$, we obtain a VARMA(1, 1) representation

$\boldsymbol{Z}_t - \boldsymbol{\Phi}_1 \boldsymbol{Z}_{t-1} = \boldsymbol{a}_t - \boldsymbol{\Theta}_1 \boldsymbol{a}_{t-1}$ in the standard VARMA form (14.4.1), but with a reduced-rank structure for the coefficient matrices such that rank $[\boldsymbol{\Phi}_1, \boldsymbol{\Theta}_1] = 1$. For a third situation, suppose the bivariate process has Kronecker indices $K_1 = 2$ and $K_2 = 1$. Then, the echelon form structure for the process is VARMA(2, 2) as in (14.7.1), with (note $K_{12} = 1$ in this case)

$$\boldsymbol{\Phi}_0^{\#} = \begin{bmatrix} 1 & 0 \\ X & 1 \end{bmatrix} \quad \boldsymbol{\Phi}_1^{\#} = \begin{bmatrix} X & 0 \\ X & X \end{bmatrix} \quad \boldsymbol{\Phi}_2^{\#} = \begin{bmatrix} X & X \\ 0 & 0 \end{bmatrix} \quad \boldsymbol{\Theta}_1^{\#} = \begin{bmatrix} X & X \\ X & X \end{bmatrix} \quad \boldsymbol{\Theta}_2^{\#} = \begin{bmatrix} X & X \\ 0 & 0 \end{bmatrix}$$

Again, on multiplication of the echelon form VARMA(2, 2) relation on the left by $\boldsymbol{\Phi}_0^{\#-1}$, we obtain a VARMA(2, 2) representation in standard form, but with reduced-rank structure for the coefficient matrices such that rank $[\boldsymbol{\Phi}_2, \boldsymbol{\Theta}_2] = 1$.

Software Implementation. In practical applications, the Kronecker index approach to model specification can be implemented using the commands Kronid, Kronfit, and refKronfit available in the MTS package of R. The specification of the Kronecker indices is performed using the command Kronid and is based on canonical correlation analysis. With the Kronecker indices specified, the VARMA parameters are estimated using the command Kronfit. Parameters with nonsignificant estimates can be removed using the command refKronfit. For further discussion and for demonstrations of the individual commands, see Tsay (2014).

14.7.2 An Empirical Example

To illustrate model specification approach described above, we return to the bivariate time series of U.S. fixed investment and change in business inventories analyzed earlier in this chapter. A bivariate VAR(2) model was fitted to the series in Section 14.2.7. As an alternative, we now consider the possibility of a mixed VARMA model for these data through determination of the echelon canonical ARMA model for the two series. The Kronecker indices $\{K_i\}$ for the process are determined using the canonical correlation method suggested by Akaike (1976) and Cooper and Wood (1982); see also Tsay (2014, Section 4.4). For the vector of present and past values, we use a maximum of three time-lagged vector variables and set $\mathbf{P}_t = (\boldsymbol{Z}_t', \boldsymbol{Z}_{t-1}', \boldsymbol{Z}_{t-2}')'$. Then, for various vectors $\mathbf{F}_{t+1}^*$ of future variables, the squared sample canonical correlations between $\mathbf{F}_{t+1}^*$ and $\mathbf{P}_t$ are determined as the eigenvalues of the matrix similar to the matrix in (6.2.6) of Section 6.2.4. The canonical correlation analysis calculations are performed sequentially by adding variables to $\mathbf{F}_{t+1}^*$ one at a time, starting with $\mathbf{F}_{t+1}^* = (z_{1,t+1})$, until $k = 2$ near zero sample canonical correlations between $\mathbf{P}_t$ and $\mathbf{F}_{t+1}^*$ are determined. At each step, a likelihood ratio test is used to determine the significance of the smallest squared canonical correlation.

The calculations can be performed using the MTS package in R. If zz denotes the two time series, the command for determining the Kronecker indices is Kronfit(zz, plag=3), where plag represents the number of elements in $\mathbf{P}_t$. The resulting squared sample canonical correlations between $\mathbf{P}_t$ and various future vectors $\mathbf{F}_{t+1}^*$ are presented in Table 14.3. From these results, we note that the first occurrence of a small squared sample canonical correlation value (0.044), indicative of a zero canonical correlation between the future and the present and past, is obtained when $\mathbf{F}_{t+1}^* = (z_{1,t+1}, z_{2,t+1}, z_{1,t+2})'$. This indicates that the Kronecker index $K_1 = 1$, since it implies that a linear combination involving $z_{1,t+2}$ in terms of the remaining variables in $\mathbf{F}_{t+1}^*$, that is, of the form $z_{1,t+2} - \boldsymbol{\phi}_1(1)' \boldsymbol{Z}_{t+1}$, is uncorrelated

TABLE 14.3 Specification of Kronecker Indices for First Differences of U.S. Fixed Investment Data and Changes in Business Inventories Data

Future Vector $\mathbf{F}^*_{t+1}$	Smallest Squared Canonical Correlation	LR Test	Degrees of Freedom	p-Value	Kronecker Index
$z_{1,t+1}$	0.371	44.02	6	0.000	
$z_{1,t+1}, z_{2,t+1}$	0.369	43.50	5	0.000	
$z_{1,t+1}, z_{2,t+1}, z_{1,t+2}$	0.044	4.13	4	0.389	$K_1 = 1$
$z_{1,t+1}, z_{2,t+1}, z_{2,t+2}$	0.069	6.20	4	0.185	$K_2 = 1$

with the present and past vector $\mathbf{P}_t$. An additional small squared canonical correlation value of 0.069 occurs when $\mathbf{F}^*_{t+1} = (z_{1,t+1}, z_{2,t+1}, z_{2,t+2})'$, and this implies that we may have $K_2 = 1$. Hence, this leads to specification of a VARMA(1, 1) model in the echelon form of equation (14.7.1) with Kronecker indices $K_1 = K_2 = 1$. This echelon model form is, in fact, the same as the standard VARMA(1, 1) model in (14.4.1); that is, $K_1 = K_2 = 1$ implies that we have $\mathbf{\Phi}^{\#}_0 = \mathbf{\Theta}^{\#}_0 = \mathbf{I}$ in (14.7.1).

The canonical correlation analysis suggests that a VARMA(1, 1) model might be essentially equivalent to the VAR(2) model in terms of fit, and that these two models are likely superior to other models considered. The parameters of the VARMA(1, 1) model were estimated using the Kronfit routine available in the MTS package of R, and the results are given as

$$\hat{\mathbf{\Phi}}_1 = \begin{bmatrix} 0.440 & -0.200 \\ (0.176) & (0.063) \\ 0.637 & 0.775 \\ (0.210) & (0.076) \end{bmatrix} \quad \hat{\mathbf{\Theta}}_1 = \begin{bmatrix} -0.030 & -0.309 \\ (0.209) & (0.081) \\ 0.313 & 0.227 \\ (0.284) & (0.129) \end{bmatrix}$$

$$\tilde{\mathbf{\Sigma}} = \begin{bmatrix} 5.0239 & 1.6697 \\ 1.6697 & 16.8671 \end{bmatrix}$$

with $|\tilde{\mathbf{\Sigma}}| = 81.9498$, and AIC = 4.608. Again, the coefficient estimate in the (1, 1) position of the matrix $\hat{\mathbf{\Theta}}_1$, as well as estimates in the second row of $\hat{\mathbf{\Theta}}_1$, is not significant and might be omitted from the model.

It is clear from these estimation results, particularly from the estimates $\tilde{\mathbf{\Sigma}}$ and associated summary measures, that the VARMA(1, 1) model provides a nearly equivalent fit to the VAR(2) model. For instance, we consider the coefficient matrices $\mathbf{\Psi}_j$ in the infinite VMA representation for $\boldsymbol{Z}_t$ implied by the VAR(2) and VARMA(1, 1) models. For the VAR(2) model, the $\mathbf{\Psi}_j$ are determined from $\mathbf{\Psi}_1 = \mathbf{\Phi}_1$:

$$\mathbf{\Psi}_j = \mathbf{\Phi}_1\mathbf{\Psi}_{j-1} + \mathbf{\Phi}_2\mathbf{\Psi}_{j-2} \quad \text{for} \quad j > 1\,(\mathbf{\Psi}_0 = \mathbf{I})$$

hence, the $\mathbf{\Psi}_j$ are given as

$$\mathbf{\Psi}_1 = \begin{bmatrix} 0.50 & 0.11 \\ 0.34 & 0.53 \end{bmatrix} \quad \mathbf{\Psi}_2 = \begin{bmatrix} 0.15 & -0.09 \\ 0.61 & 0.46 \end{bmatrix} \quad \mathbf{\Psi}_3 = \begin{bmatrix} -0.00 & -0.12 \\ 0.55 & 0.31 \end{bmatrix}$$

$$\mathbf{\Psi}_4 = \begin{bmatrix} -0.09 & -0.11 \\ 0.41 & 0.16 \end{bmatrix} \quad \mathbf{\Psi}_5 = \begin{bmatrix} -0.11 & -0.08 \\ 0.26 & 0.06 \end{bmatrix} \quad \mathbf{\Psi}_6 = \begin{bmatrix} -0.10 & -0.05 \\ 0.14 & -0.00 \end{bmatrix}$$

and so on, while those for the VARMA(1, 1) model are determined from $\boldsymbol{\Psi}_1 = \boldsymbol{\Phi}_1 - \boldsymbol{\Theta}_1, \boldsymbol{\Psi}_j = \boldsymbol{\Phi}_1 \boldsymbol{\Psi}_{j-1}, j > 1$, and so are given as

$$\boldsymbol{\Psi}_1 = \begin{bmatrix} 0.47 & 0.11 \\ 0.32 & 0.55 \end{bmatrix} \quad \boldsymbol{\Psi}_2 = \begin{bmatrix} 0.14 & -0.06 \\ 0.55 & 0.49 \end{bmatrix} \quad \boldsymbol{\Psi}_3 = \begin{bmatrix} -0.05 & -0.13 \\ 0.52 & 0.34 \end{bmatrix}$$
$$\boldsymbol{\Psi}_4 = \begin{bmatrix} -0.12 & -0.12 \\ 0.37 & 0.19 \end{bmatrix} \quad \boldsymbol{\Psi}_5 = \begin{bmatrix} -0.13 & -0.09 \\ 0.21 & 0.07 \end{bmatrix} \quad \boldsymbol{\Psi}_6 = \begin{bmatrix} -0.10 & -0.05 \\ 0.08 & -0.01 \end{bmatrix}$$

Thus, we see that the $\boldsymbol{\Psi}_j$ coefficient matrices are very similar for both models, implying, in particular, that forecasts $\hat{\boldsymbol{Z}}_t(l)$ and the covariance matrices $\boldsymbol{\Sigma}(l) + \sum_{j=0}^{l-1} \boldsymbol{\Psi}_j \boldsymbol{\Sigma} \boldsymbol{\Psi}_j'$ of the l-step-ahead forecast errors $\boldsymbol{e}_t(l) = \boldsymbol{Z}_{t+l} - \hat{\boldsymbol{Z}}_t(l)$ obtained from the two models, VAR(2) and VARMA(1, 1), are nearly identical.

14.7.3 Partial Canonical Correlation Analysis for Reduced-Rank Structure

Another approach to allow for simplifying structure in the parameterization of the VAR and VARMA models is to incorporate certain *reduced-rank* structure in the coefficient matrices. For the VAR(p) model (14.2.1), Ahn and Reinsel (1988) proposed a particular nested reduced-rank model structure, such that

$$\text{rank}(\boldsymbol{\Phi}_j) = r_j \geq \text{rank}\,(\boldsymbol{\Phi}_{j+1}) = r_{j+1} \quad j = 1, 2, \ldots, p-1$$

and it is also specified that range($\boldsymbol{\Phi}_j$) $\supset$ range($\boldsymbol{\Phi}_{j+1}$). Then the $\boldsymbol{\Phi}_j$ can be represented in reduced-rank factorization form as $\boldsymbol{\Phi}_j = \mathbf{A}_j \mathbf{B}_j$, where $\mathbf{A}_j$ and $\mathbf{B}_j$ are full-rank matrices of dimensions $k \times r_j$ and $r_j \times k$, respectively, with range($\mathbf{A}_j$) $\supset$ range($\mathbf{A}_{j+1}$). One fundamental consequence for this model is that there then exists a full-rank $(k - r_j) \times k$ matrix $\mathbf{F}_j'$, such that $\mathbf{F}_j' \boldsymbol{\Phi}_j = \mathbf{0}$ and hence $\mathbf{F}_j' \boldsymbol{\Phi}_i = \mathbf{0}$ for all $i \geq j$ because of the nested structure. Therefore, the vector

$$\mathbf{F}_j' \left(\boldsymbol{Z}_t - \sum_{i=1}^{j-1} \boldsymbol{\Phi}_i \boldsymbol{Z}_{t-i} \right) = \mathbf{F}_j' \left(\boldsymbol{Z}_t - \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{Z}_{t-i} \right) \equiv \mathbf{F}_j' \boldsymbol{\delta} + \mathbf{F}_j' \boldsymbol{a}_t$$

is uncorrelated with the past values $\boldsymbol{Z}_{j-1,t-1} = (\boldsymbol{Z}_{t-1}', \ldots, \boldsymbol{Z}_{t-j}')'$ and consists of $k - r_j$ linear combinations of $\boldsymbol{Z}_{j-1,t} = (\boldsymbol{Z}_t', \ldots, \boldsymbol{Z}_{t-j+1}')'$. Thus, it follows that $k - r_j$ *zero partial canonical correlations* will occur between $\boldsymbol{Z}_t$ and $\boldsymbol{Z}_{t-j}$, given $\boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_{t-j+1}$. Hence, performing a (partial) canonical correlation analysis for the various values of $j = 1, 2, \ldots$ can identify the simplifying nested reduced-rank structure, as well as the overall order p, of the VAR model.

The sample statistic that can be used to (tentatively) specify the ranks is

$$C(j, r) = -(N - j - jk - 1) \sum_{t=r+1}^{k} \ln[1 - \hat{\rho}_i^2(j)] \tag{14.7.8}$$

for $r = k - 1, k - 2, \ldots, 0$, where $1 \geq \hat{\rho}_1(j) \geq \cdots \geq \hat{\rho}_k(j) > 0$ are the *sample partial canonical correlations* between $\boldsymbol{Z}_t$ and $\boldsymbol{Z}_{t-j}$, given $\boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_{t-j+1}$. (Calculation of sample canonical correlations was discussed previously in Section 6.2.4.) Under the null hypothesis that rank($\boldsymbol{\Phi}_j$) $\leq r$ within the nested reduced-rank model framework, the statistic $C(j, r)$ is

asymptotically distributed as chi-squared with $(k-r)^2$ degrees of freedom. Hence, if the value of the test statistic is not ''significantly'' large, we would not reject the null hypothesis and might conclude that $\boldsymbol{\Phi}_j$ has reduced rank equal to the smallest value r_j for which the test does not reject the null hypothesis. Note, in particular, that when $r = 0$ the statistic in (14.7.8) is (essentially) the same as the LR test statistic given in (14.2.10) for testing $H_0\colon \boldsymbol{\Phi}_j = \mathbf{0}$ in an VAR(j) model, since it can be verified that $\ln[|\mathbf{S}_j|/|\mathbf{S}_{j-1}|] = \sum_{i=1}^{k} \ln[1-\hat{\rho}_i^2(j)]$.

Once the ranks in the nested reduced-rank VAR model have been specified, the parameters in the restricted model can be estimated by maximum likelihood methods. Some normalization conditions on the $\mathbf{A}_j$ and $\mathbf{B}_j$ in $\boldsymbol{\Phi}_j = \mathbf{A}_j\mathbf{B}_j$ are required to ensure a unique set of parameters. Assuming the components of $\boldsymbol{Z}_t$ are arranged suitably, this parameterization can be obtained as $\boldsymbol{\Phi}_j = \mathbf{A}_1\mathbf{D}_j\mathbf{B}_j$, where $\mathbf{A}_1$ is $k \times r_1$ lower triangular with ones on the main diagonal and may have certain other elements ''normalized'' to fixed values of zero, $\mathbf{B}_j$ contains unrestricted parameters, and $\mathbf{D}_j = [\mathbf{I}_{r_j}, \mathbf{0}]'$ is $r_1 \times r_j$. Asymptotic distribution theory for the ML estimators of parameters of this model extends from theory for the LS estimators in a stationary VAR(p) model in a fairly direct manner.

The structure of the reduced-rank VAR model relates directly to the concepts of Kronecker indices, McMillan degree, and echelon canonical form of VARMA models discussed earlier. In particular, it can be easily verified that the McMillan degree of a nested reduced-rank AR process is equal to $M = \sum_{j=1}^{p} r_j$, the sum of the ranks of the AR coefficient matrices $\boldsymbol{\Phi}_j$. In addition, from the nested reduced-rank structure it follows that the model can also be represented as

$$\boldsymbol{\Phi}_0^{\#}\boldsymbol{Z}_t - \sum_{j=1}^{p} \boldsymbol{\Phi}_j^{\#}\boldsymbol{Z}_{t-j} = \boldsymbol{\delta}^{\#} + \boldsymbol{\Phi}_0^{\#}\boldsymbol{a}_t$$

with $\boldsymbol{\Phi}_0^{\#} = \mathbf{A}^{-1}$, where $\mathbf{A}$ is the $k \times k$ matrix formed by augmenting the $k \times r_1$ matrix $\mathbf{A}_1$ with the last $k - r_1$ columns of the $k \times k$ identity matrix, and

$$\boldsymbol{\Phi}_j^{\#} = \mathbf{A}^{-1}\boldsymbol{\Phi}_j = \mathbf{A}^{-1}\mathbf{A}_1\mathbf{D}_j\mathbf{B}_j \equiv [\mathbf{B}_j', \mathbf{0}']'$$

having its last $k - r_j$ rows equal to zero. This relation can be viewed as an echelon canonical form representation, as in (14.7.1), for the nested reduced-rank vector VAR(p) model. Also, as noted by Reinsel (1997, p. 66), the notion of a nested reduced-rank model and its relationship to the echelon form representation can be directly extended to the VARMA model leading to the specification of a reduced-rank VARMA model for the vector process.

14.8 NONSTATIONARITY AND COINTEGRATION

14.8.1 Vector ARIMA Models

Time series encountered in practice will frequently exhibit nonstationary behavior. To generalize stationary VARMA models to nonstationary processes, we can consider a general form of the VARMA model, $\boldsymbol{\Phi}(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}(B)\mathbf{a}_t$, where some of the roots of $\det\{\boldsymbol{\Phi}(B)\} = 0$ are allowed to have absolute value equal to one. More specifically, because of the prominent role of the differencing operator $(1 - B)$ in univariate models, for nonseasonal time series we might only allow some roots to equal one (unit roots) while the remaining roots are all

greater than one in absolute value. A particular restrictive class of models of this type for nonstationary series are of the form

$$\boldsymbol{\Phi}_1(B)\boldsymbol{D}(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}(B)\mathbf{a}_t \tag{14.8.1}$$

where $\mathbf{D}(B) = \text{diag}[(1-B)^{d_1}, \ldots, (1-B)^{d_k}]$ is a diagonal matrix, $d_1, \ldots, d_k$ are nonnegative integers, and $\det\{\boldsymbol{\Phi}_1(B)\} = 0$ has all roots greater than one in absolute value. Thus, this model, which is referred to as a vector ARIMA model, simply states that after each series z_{it} is individually differenced an appropriate number (d_i) of times to reduce it to a stationary series, the resulting vector series $\mathbf{W}_t = \mathbf{D}(B)\boldsymbol{Z}_t$ is a stationary VARMA(p, q) process. For vector time series, however, simultaneous differencing of all component series can lead to unnecessary complications in modeling and estimation as a result of ''overdifferencing,'' including noninvertible model representations, so differencing needs to be examined with particular care in the vector case.

14.8.2 Cointegration in Nonstationary Vector Processes

The nonstationary unit-root aspects of a vector process $\boldsymbol{Z}_t$ become more complicated in the multivariate case compared with the univariate case, due in part to the possibility of *cointegration* among the component series z_{it} of a nonstationary vector process $\boldsymbol{Z}_t$. For instance, the possibility exists for each component series z_{it} to be nonstationary with its first difference $(1-B)z_{it}$ stationary (in which case z_{it} is said to be integrated of order one), but such that certain linear combinations $y_{it} = \boldsymbol{b}_i'\boldsymbol{Z}_t$ of $\boldsymbol{Z}_t$ will be stationary. That this possibility exists was demonstrated by Box and Tiao (1977) in their analysis of a five-dimensional dataset from Quenouille (1957). A process $\boldsymbol{Z}_t$ that displays this behavior is said to be *cointegrated* with cointegrating vectors $\boldsymbol{b}_i$ (e.g., Engle and Granger, 1987). An interpretation of cointegrated vector processes $\boldsymbol{Z}_t$ is that the individual components z_{it} share some common nonstationary components or ''common trends''; hence, they tend to have certain similar movements in their longer term behavior. These common trend components will be eliminated upon taking suitable linear combinations of the components of the process $\boldsymbol{Z}_t$. A related interpretation is that the component series z_{it}, although they may exhibit nonstationary behavior, satisfy a long-run equilibrium relation $\boldsymbol{b}_i'\boldsymbol{Z}_t \simeq 0$ such that the process $y_{it} = \boldsymbol{b}'_i\boldsymbol{Z}_t$, which represents the deviation from the equilibrium, exhibits stable behavior and so forms a stationary process. Properties of nonstationary cointegrated systems have been investigated by Engle and Granger (1987) and Johansen (1988), among others.

An Error Correction Form. A specific nonstationary VARMA model structure for which cointegration occurs is the model $\boldsymbol{\Phi}(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}(B)\boldsymbol{a}_t$, where $\det\{\boldsymbol{\Phi}(B)\} = 0$ has $d < k$ roots equal to one and all other roots are greater than one in absolute value, and also the matrix

$$\boldsymbol{\Phi}(1) = \mathbf{I} - \boldsymbol{\Phi}_1 - \cdots - \boldsymbol{\Phi}_p$$

has rank $r = k - d$. Because the process has unit roots fewer than the number of components, this type of process is called *partially nonstationary* by Ahn and Reinsel (1990). For such a process, it can be established that r linearly independent vectors $\boldsymbol{b}_i$ exist such that $\boldsymbol{b}_i'\boldsymbol{Z}_t$ is stationary, and $\boldsymbol{Z}_t$ is said to have cointegrating rank r. A useful approach to the investigation of this model is to express it in its equivalent *error correction* (EC) form

given by

$$\boldsymbol{W}_t = \mathbf{C}\boldsymbol{Z}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Phi}_j^* \boldsymbol{W}_{t-j} + \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j} \tag{14.8.2}$$

where $\boldsymbol{W}_t = (1 - B)\boldsymbol{Z}_t, \boldsymbol{\Phi}_j^* = -\sum_{i=j+1}^{p} \boldsymbol{\Phi}_i$, and

$$\mathbf{C} = -\boldsymbol{\Phi}(1) = -\left(\boldsymbol{I} - \sum_{j=1}^{p} \boldsymbol{\Phi}_j\right) \tag{14.8.3}$$

For instance, by subtracting $\boldsymbol{Z}_{t-1}$ from both sides of the VAR(1) model $\boldsymbol{Z}_t = \boldsymbol{\Phi}\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t$, we see that the model can be expressed as $(\boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}) = -(\mathbf{I} - \boldsymbol{\Phi})\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t \equiv \mathbf{C}\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t$, with $\mathbf{C} = -(\mathbf{I} - \boldsymbol{\Phi})$. The VAR(2) model can be expressed as

$$\begin{aligned}(\boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}) &= -(\mathbf{I} - \boldsymbol{\Phi}_1 - \boldsymbol{\Phi}_2)\boldsymbol{Z}_{t-1} - \boldsymbol{\Phi}_2(\boldsymbol{Z}_{t-1} - \boldsymbol{Z}_{t-2}) + \boldsymbol{a}_t \\ &\equiv \mathbf{C}\boldsymbol{Z}_{t-1} + \boldsymbol{\Phi}_1^*(\boldsymbol{Z}_{t-1} - \boldsymbol{Z}_{t-2}) + \boldsymbol{a}_t\end{aligned}$$

with $\mathbf{C} = -(\mathbf{I} - \boldsymbol{\Phi}_1 - \boldsymbol{\Phi}_2)$ and $\boldsymbol{\Phi}_1^* = -\boldsymbol{\Phi}_2$, and similarly for higher order VAR models.

We note that the error correction form (14.8.2) has an invertible moving average operator but introduces $\mathbf{C}\boldsymbol{Z}_{t-1}$ on the right-hand side of the model. Since the moving average operator remains unchanged, problems associated with noninvertibility are now avoided. The term $\mathbf{C}\boldsymbol{Z}_{t-1}$ is referred to as the error correction term and the rank $r = k - d$ of the coefficient matrix $\mathbf{C}$ represents the number of cointegrating vectors in the system.

To derive an alternative form, we note that the reduced-rank matrix $\mathbf{C}$ can be written as $\mathbf{C} = \mathbf{AB}$, where $\mathbf{A}$ and $\mathbf{B}$ are full-rank matrices of dimensions $k \times r$ and $r \times k$, respectively. We can also determine a full-rank $k \times (k - r)$ matrix $\mathbf{Q}_1$ such that $\mathbf{Q}_1'\mathbf{A} = \mathbf{0}$, hence also $\mathbf{Q}_1'\mathbf{C} = \mathbf{0}$. Hence, it can be established that the r linear combinations $\boldsymbol{Y}_{2t} = \mathbf{B}\boldsymbol{Z}_t$ are stationary, the r rows of $\mathbf{B}$ are linearly independent cointegrating vectors, whereas the $d = k - r$ components $\boldsymbol{Y}_{1t} = \mathbf{Q}'_1\boldsymbol{Z}_t$ are ''purely'' nonstationary and are often referred to as the ''common trends'' among the components of the nonstationary process $\boldsymbol{Z}_t$. Therefore, the error correction form (14.8.2) can also be expressed as

$$\begin{aligned}\boldsymbol{W}_t &= \mathbf{AB}\mathbf{Z}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Phi}_j^* \boldsymbol{W}_{t-j} + \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j} \\ &\equiv \mathbf{A}\boldsymbol{Y}_{2,t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Phi}_j^* \boldsymbol{W}_{t-j} + \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j}\end{aligned} \tag{14.8.4}$$

Issues of estimation of cointegrated VAR models and testing for the rank r of cointegration will be discussed briefly in Section 14.8.3.

Illustration: Nonstationary VAR(1) Model. To illustrate some of the preceding points, consider the VAR(1) process $\boldsymbol{Z}_t = \boldsymbol{\Phi}\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t$ with d eigenvalues of $\boldsymbol{\Phi}$ equal to one and the remaining $r = k - d$ eigenvalues less than one in absolute value, and suppose the d unit eigenvalues have d linearly independent eigenvectors. Then there is a $k \times k$ nonsingular

matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{\Phi}\mathbf{P} = \mathbf{\Lambda}$ with $\mathbf{\Lambda} = \text{diag}(\mathbf{I}_d, \mathbf{\Lambda}_2)$, where $\mathbf{\Lambda}_2 = \text{diag}(\lambda_{d+1}, \ldots, \lambda_k)$ is an $r \times r$ diagonal matrix with $|\lambda_i| < 1$. Letting $\mathbf{P} = [\mathbf{P}_1, \mathbf{P}_2]$ and $\mathbf{Q} = \mathbf{P}^{-1} = [\mathbf{Q}_1, \mathbf{Q}_2]'$, where $\mathbf{P}_1$ and $\mathbf{Q}_1$ are $k \times d$ matrices, define $\boldsymbol{Y}_t = \mathbf{Q}\boldsymbol{Z}_t = (\boldsymbol{Y}'_{1t}, \boldsymbol{Y}'_{2t})'$, that is, $\boldsymbol{Y}_{1t} = \mathbf{Q}'_1\boldsymbol{Z}_t$ and $\boldsymbol{Y}_{2t} = \mathbf{Q}'_2\boldsymbol{Z}_t$, and similarly $\boldsymbol{\varepsilon}_t = \mathbf{Q}\boldsymbol{a}_t = (\boldsymbol{\varepsilon}'_{1t}, \boldsymbol{\varepsilon}'_{2t})'$. Then we have

$$\mathbf{Q}\boldsymbol{Z}_t = \mathbf{Q}\mathbf{\Phi}\mathbf{P}\mathbf{Q}\boldsymbol{Z}_{t-1} + \mathbf{Q}\boldsymbol{a}_t$$

or $\boldsymbol{Y}_t = \mathbf{\Lambda}\boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t$. Therefore, the model in terms of $\boldsymbol{Y}_t$ reduces to

$$(1 - B)\boldsymbol{Y}_{1t} = \boldsymbol{\varepsilon}_{1t} \quad \text{and} \quad (\mathbf{I} - \mathbf{\Lambda}_2 B)\boldsymbol{Y}_{2t} = \boldsymbol{\varepsilon}_{2t}$$

so $\{\boldsymbol{Y}_{1t}\}$ is a d-dimensional purely nonstationary series, whereas $\{\boldsymbol{Y}_{2t}\}$ is an r-dimensional stationary series. Thus, $\{\boldsymbol{Z}_t\}$ is nonstationary but has r linearly independent linear combinations $\boldsymbol{Y}_{2t} = \mathbf{Q}'_2\boldsymbol{Z}_t$, which are stationary, so $\boldsymbol{Z}_t$ is cointegrated with cointegrating rank r and linearly independent cointegrating vectors, which are the rows of $\mathbf{Q}'_2$. Conversely, since $\boldsymbol{Z}_t = \mathbf{P}\boldsymbol{Y}_t = \mathbf{P}_1\boldsymbol{Y}_{1t} + \mathbf{P}_2\boldsymbol{Y}_{2t}$, the components of the vector $\boldsymbol{Z}_t$ are linear combinations of a nonstationary vector (random walk) component $\boldsymbol{Y}_{1t}$ and a stationary VAR(1) component $\boldsymbol{Y}_{2t}$. Also notice that the error correction form of this VAR(1) model as in (14.8.2) is

$$\boldsymbol{W}_t = \boldsymbol{Z}_t - \boldsymbol{Z}_{t-1} = \mathbf{C}\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t$$

where

$$\mathbf{C} = -(\mathbf{I} - \mathbf{\Phi}) = -\mathbf{P}(\mathbf{I} - \mathbf{\Lambda})\mathbf{Q} = -\mathbf{P}_2(\mathbf{I}_r - \mathbf{\Lambda}_2)\mathbf{Q}'_2$$

which is clearly of reduced rank r.

14.8.3 Estimation and Inferences for Cointegrated VAR Models

As noted above, when the vector series $\boldsymbol{Z}_t$ is unit-root nonstationary but with cointegration features, it is not appropriate to difference all component series and model the resulting series $\mathbf{W}_t = (1 - B)\boldsymbol{Z}_t$ by a VAR or VARMA model. Instead, we may prefer to incorporate the unit-root and cointegration features into the analysis using the model (14.8.2). This can provide a better understanding on the nature of the nonstationarity and improve the forecasting performance of the model. This section examines the estimation and statistical inference for this model focusing on the special case of an error correction VAR(p) model

$$\boldsymbol{W}_t = \mathbf{C}\boldsymbol{Z}_{t-1} + \sum_{j=1}^{p-1} \mathbf{\Phi}^*_j \boldsymbol{W}_{t-j} + \boldsymbol{a}_t \tag{14.8.5}$$

where $\boldsymbol{W}_t = \boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}$ with $\text{rank}(\mathbf{C}) = r < k$. Note that a *special case* of (14.8.5), at one extreme, occurs with $r = 0$ (i.e., $d = k$ unit roots and $\mathbf{C} = \mathbf{0}$) and leads to a usual VAR model of order $p - 1$ for the series of first differences $\boldsymbol{W}_t$.

The least-squares and Gaussian maximum likelihood estimation of cointegrated VAR models and likelihood ratio testing for the rank of cointegration, generally utilizing the error correction form (14.8.5) of the model, have been examined by several authors including Johansen (1988, 1991), Johansen and Juselius (1990), Ahn and Reinsel (1990), and Reinsel and Ahn (1992). Estimation of the cointegrated model, which imposes the restriction on the number d of unit roots in $\mathbf{\Phi}(B)$ (or the number $r = k - d$ of cointegrating relations), is equivalent to *reduced-rank* estimation, which imposes the restriction on the rank r of

the coefficient matrix $\mathbf{C}$, which can be written as $\mathbf{C} = \mathbf{AB}$ as noted in Section 14.8.2. So techniques from reduced-rank estimation of multivariate regression models can be utilized.

When there are no additional constraints on the coefficient matrices $\mathbf{\Phi}_j^*$ in (14.8.5), that is, the only parameter constraints involved in the model are $\text{rank}(\mathbf{C}) \leq r$, it follows from the original work by Anderson (1951) on reduced-rank regression that the Gaussian (ML) reduced-rank estimation can be obtained explicitly through the partial canonical correlation analysis between $\boldsymbol{W}_t$ and $\boldsymbol{Z}_{t-1}$, given $\boldsymbol{W}_{t-1}, \ldots, \boldsymbol{W}_{t-p+1}$. When there are additional constraints, however, iterative numerical techniques are needed for the Gaussian ML estimation. Specifically, in the partial canonical correlation analysis approach, let $\tilde{\boldsymbol{W}}_t$ and $\tilde{\boldsymbol{Z}}_{t-1}$ denote the residual vectors from least-squares regressions of $\boldsymbol{W}_t$ and $\boldsymbol{Z}_{t-1}$, respectively, on the lagged values $\boldsymbol{W}_{t-1}, \ldots, \boldsymbol{W}_{t-p+1}$, and let

$$\mathbf{S}_{\tilde{w}\tilde{w}} = \sum_{t=1}^{N} \tilde{\boldsymbol{W}}_t \tilde{\boldsymbol{W}}_t' \quad \mathbf{S}_{\tilde{w}\tilde{z}} = \sum_{t=1}^{N} \tilde{\boldsymbol{W}}_t \tilde{\boldsymbol{Z}}_{t-1}' \quad \mathbf{S}_{\tilde{z}\tilde{z}} = \sum_{t=1}^{N} \tilde{\boldsymbol{Z}}_{t-1} \tilde{\boldsymbol{Z}}_{t-1}'$$

Then the *Gaussian reduced-rank estimator* of $\mathbf{C}$ in model (14.8.5) can be expressed explicitly as

$$\tilde{\mathbf{C}} = \hat{\mathbf{\Sigma}}\hat{\mathbf{V}}\hat{\mathbf{V}}'\hat{\mathbf{C}} \tag{14.8.6}$$

where $\hat{\mathbf{C}} = \mathbf{S}_{\tilde{w}\tilde{z}}\mathbf{S}_{\tilde{z}\tilde{z}}^{-1}$ is the full-rank LS estimator of $\mathbf{C}$, $\hat{\mathbf{\Sigma}}$ is the corresponding residual covariance matrix estimate of $\mathbf{\Sigma} = \text{cov}[\mathbf{a}_t]$ from the full-rank LS estimation of (14.8.5), and $\hat{\mathbf{V}} = [\hat{\boldsymbol{V}}_1, \ldots, \hat{\boldsymbol{V}}_r]$ are the vectors corresponding to the r largest partial canonical correlations $\hat{\rho}_i(p), i = 1, \ldots, r$. The vectors $\hat{\boldsymbol{V}}_i$ are normalized so that $\hat{\mathbf{V}}'\hat{\mathbf{\Sigma}}\hat{\mathbf{V}} = \mathbf{I}_r$. Note that the form of the estimator (14.8.6) provides the reduced-rank factorization as $\tilde{\mathbf{C}} = (\hat{\mathbf{\Sigma}}\hat{\mathbf{V}})(\hat{\mathbf{V}}'\hat{\mathbf{C}}) \equiv \hat{\mathbf{A}}\hat{\mathbf{B}}$, with $\hat{\mathbf{A}} = \hat{\mathbf{\Sigma}}\hat{\mathbf{V}}$ satisfying the normalization $\hat{\mathbf{A}}'\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{A}} = \mathbf{I}_r$.

The asymptotic distribution theory of the LS and reduced-rank estimators, $\hat{\mathbf{C}}$ and $\tilde{\mathbf{C}}$, and of LR test statistics for rank has been established and the limiting distributions represented as functionals of vector Brownian motion processes, extending the ''nonstandard'' unit-root asymptotic distribution theory for univariate AR models as outlined in Section 10.1. In particular, we discuss the LR statistic for the test of the hypothesis H_0: $\text{rank}(\mathbf{C}) \leq r$ for model (14.8.5). The LR test statistic is given by

$$-N \ln(U) = -N \ln\left(\frac{|\mathbf{S}|}{|\mathbf{S}_0|}\right)$$

where $\mathbf{S}$ denotes the residual sum-of-squares matrix in the full-rank LS estimation (such that $\hat{\mathbf{\Sigma}} = N^{-1}\mathbf{S}$), while $\mathbf{S}_0$ is the residual sum-of-squares matrix obtained under the reduced-rank restriction that $\text{rank}(\mathbf{C}) = r$. Again from the work of Anderson (1951), it is established that

$$\mathbf{S}_0 = \mathbf{S} + (\hat{\mathbf{C}} - \tilde{\mathbf{C}})\mathbf{S}_{\tilde{z}\tilde{z}}(\hat{\mathbf{C}} - \tilde{\mathbf{C}})'$$

and it follows that

$$|\mathbf{S}_0| = |\mathbf{S}| \prod_{i=r+1}^{k} [1 - \hat{\rho}_i^2(p)]^{-1}$$

where the $\hat{\rho}_i(p)$ are the $d = k - r$ *smallest* sample partial canonical correlations between $\boldsymbol{W}_t$ and $\boldsymbol{Z}_{t-1}$, given $\boldsymbol{W}_{t-1} \ldots, \boldsymbol{W}_{t-p+1}$. Therefore, the LR statistic can be expressed equivalently as

$$-N \ln(U) = -N \sum_{i=r+1}^{k} \ln[1 - \hat{\rho}_i^2(p)] \tag{14.8.7}$$

The limiting distribution for the LR statistic has been derived, based on limiting distribution properties for the LS and reduced-rank estimators $\hat{\mathbf{C}}$ and $\tilde{\mathbf{C}}$, and its limiting distribution is represented by

$$-N \ln(U) \xrightarrow{\mathcal{D}} \operatorname{tr}\left\{ \left[\int_0^1 \mathbf{B}_d(u) d\mathbf{B}_d(u)' \right]' \left[\int_0^1 \mathbf{B}_d(u)\mathbf{B}_d(u)' du \right]^{-1} \times \left[\int_0^1 \mathbf{B}_d(u) d\mathbf{B}_d(u)' \right] \right\} \tag{14.8.8}$$

where $\mathbf{B}_d(u)$ is a d-dimensional standard Brownian motion process, with $d = k - r$. The limiting distribution of the LR statistic under H_0 depends only on d and not on any nuisance parameters or the order p of the VAR model. Note that in the special case of testing for (at least) one unit root, $d = 1$, the limiting distribution in (14.8.8) reduces to

$$-N \ln(U) \xrightarrow{\mathcal{D}} \frac{\left[\int_0^1 B_1(u) d B_1(u) \right]^2}{\int_0^1 B_1(u)^2 du}$$

which is the asymptotic distribution for the (univariate) unit root statistic $\hat{\tau}^2$ in the univariate AR(1) model as discussed in Section 10.1.1.

Critical values of the limiting distribution in (14.8.8) have been obtained by simulation by Johansen (1988) and Reinsel and Ahn (1992) and can be used in the test of H_0. Similar to other LR testing procedures in multivariate linear models, it is suggested that the LR statistic in (14.8.7) be modified to $-(N - kp) \sum_{i=r+1}^{k} \ln[1 - \hat{\rho}_i^2(p)]$ for practical use in finite samples, as this may provide a test statistic whose finite sample distribution is closer to the limiting distribution in (14.8.8) than the ''unmodified'' LR test statistic. The ML estimation and LR testing procedures and asymptotic theory are also extended to the more practical case where a constant term $\boldsymbol{\delta}$ is included in the estimation of the VAR(p) model in error correction form, $\boldsymbol{W}_t = \mathbf{C}\boldsymbol{Z}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Phi}_j^* \boldsymbol{W}_{t-j} + \boldsymbol{\delta} + \mathbf{a}_t$. A recommended procedure to be used in specification of the rank r or $\mathbf{C}$ in model (14.8.5) is thus based on performing LR tests of H_0: rank$(\mathbf{C}) \le r$ for a sequence of values of $r = k - 1, k - 2, \ldots, 1, 0$, and an appropriate value of r can be chosen as the smallest value for which H_0 is not rejected. For further discussion of the model building process and for software demonstrations using the R package, the readers are referred to Tsay (2014).

APPENDIX A14.1 SPECTRAL CHARACTERISTICS AND LINEAR FILTERING RELATIONS FOR STATIONARY MULTIVARIATE PROCESSES

A14.1.1 Spectral Characteristics for Stationary Multivariate Processes

The covariance-generating matrix function (provided $\sum_{l=-\infty}^{\infty} |\gamma_{i,j}(l)| < \infty, i, j = 1, \ldots, k$) is defined as $\mathbf{G}(z) = \sum_{l=-\infty}^{\infty} \mathbf{\Gamma}(l) z^l$, and the *spectral density* matrix of the stationary process $\{\boldsymbol{Z}_t\}$ as a function of frequency f is defined as

$$\mathbf{P}(f) = 2\mathbf{G}(e^{-i2\pi f}) = 2 \sum_{l=-\infty}^{\infty} \mathbf{\Gamma}(l) e^{-i2\pi f l} \qquad 0 \le f \le \frac{1}{2} \tag{A14.1.1}$$

The (h, j)th element of $\mathbf{P}(f)$, denoted as $p_{hj}(f)$, is

$$p_{hj}(f) = 2 \sum_{l=-\infty}^{\infty} \gamma_{hj}(l) e^{-i2\pi f l}$$

For $h = j, p_{jj}(f)$ is the (auto)spectral density function of the series z_{jt}, while for $h \neq j$, $p_{hj}(f)$ is the cross-spectral density function of z_{ht} and z_{jt}. Notice that $p_{jj}(f)$ is real valued and nonnegative, but since $\gamma_{hj}(l) \neq \gamma_{hj}(-l)$ for $h \neq j$, the cross-spectral density function $p_{hj}(f)$ is in general complex valued, with $p_{hj}(f)$ being equal to $p_{jh}(-f)$, the complex conjugate of $p_{jh}(f)$. Therefore, the spectral density matrix $\mathbf{P}(f)$ is Hermitian, that is, $\mathbf{P}(f) = \mathbf{P}(-f)'$. Moreover, $\mathbf{P}(f)$ is a nonnegative-definite matrix in the sense that $\boldsymbol{b}'\mathbf{P}(f)\boldsymbol{b} \ge 0$ for any k-dimensional (real-valued) vector $\boldsymbol{b}$, since $\boldsymbol{b}'\mathbf{P}(f)\boldsymbol{b}$ is the spectral density function of the linear combination $\boldsymbol{b}'\boldsymbol{Z}_t$ and hence must be nonnegative. Note also that

$$\mathbf{\Gamma}(l) = \frac{1}{2} \int_{-1/2}^{1/2} e^{i2\pi f l} \mathbf{P}(f)\, df \quad l = 0, \pm 1, \pm 2, \ldots \tag{A14.1.2}$$

that is, $\gamma_{hj}(l) = \frac{1}{2} \int_{-1/2}^{1/2} e^{i2\pi f l} p_{hj}(f)\, df$.

The real part of $p_{hj}(f)$, denoted as $c_{hj}(f) = \mathrm{Re}\{p_{hj}(f)\}$, is called the *co-spectrum*, and the negative of the imaginary part, denoted as $q_{hj}(f) = -\mathrm{Im}\{p_{hj}(f)\}$, is called the *quadrature spectrum*. We can also express $p_{hj}(f)$ in polar form as $p_{hj}(f) = \alpha_{hj}(f) e^{i\phi_{hj}(f)}$, where

$$\alpha_{hj}(f) = |p_{hj}(f)| = \{c_{hj}^2(f) + q_{hj}^2(f)\}^{1/2}$$

and $\phi_{hj}(f) = \tan^{-1}\{-q_{hj}(f)/c_{hj}(f)\}$. The function $\alpha_{hj}(f)$ is called the *cross-amplitude spectrum* and $\phi_{hj}(f)$ is the *phase spectrum*.

Similar to the univariate case, the spectral density matrix $\mathbf{P}(f)$ represents the covariance matrix of the random vector of components at frequency f in the theoretical spectral representations of the components z_{jt} of the vector process $\{\boldsymbol{Z}_t\}$ corresponding to the finite-sample Fourier representations of the time series z_{jt} as in (2.2.1). The (*squared*) coherency spectrum of a pair of series z_{ht} and z_{jt} is defined as

$$k_{hj}^2(f) = \frac{|p_{hj}(f)|^2}{\{p_{hh}(f) p_{jj}(f)\}}$$

The coherency $k_{hj}(f)$ at frequency f can thus be interpreted as the correlation coefficient between the random components at frequency f in the theoretical spectral representations of z_{ht} and z_{jt}. Hence, $k_{hj}(f)$ as a function of f measures the extent to which the two processes z_{ht} and z_{jt} are linearly related in terms of the degree of linear association of their random components at different frequencies f. When spectral relations that involve more than two time series are considered, the related concepts of partial coherency and multiple coherency are also of interest. Detailed accounts of the spectral theory and analysis of multivariate time series may be found in the books by Jenkins and Watts (1968), Hannan (1970), Priestley (1981), and Bloomfield (2000).

A14.1.2 Linear Filtering Relations for Stationary Multivariate Processes

The representation of dynamic linear relationships through the formulation of linear filters is fundamental to the study of stationary multivariate time series. An important example is the moving average representation of the k-dimensional process $\boldsymbol{Z}_t$ in (14.1.4). More generally, a multivariate *linear* (*time-invariant*) *filter* relating an r-dimensional input series $\boldsymbol{X}_t$ to a k-dimensional output series $\boldsymbol{Z}_t$ is given by the form

$$\boldsymbol{Z}_t = \sum_{j=-\infty}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{X}_{t-j} \tag{A14.1.3}$$

where the $\boldsymbol{\Psi}_j$ are $k \times r$ matrices. The filter is physically realizable or *causal* when the $\boldsymbol{\Psi}_j = 0$ for $j < 0$, so that $\boldsymbol{Z}_t = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j X_{t-j}$ is expressible in terms of only present and past values of the input process $\{\boldsymbol{X}_t\}$. The filter is said to be *stable* if $\sum_{j=-\infty}^{\infty} \|\boldsymbol{\Psi}_j\| < \infty$, where $\|\mathbf{A}\|$ denotes a norm for the matrix $\mathbf{A}$ such as $\|\mathbf{A}\|^2 = \text{tr}\{\mathbf{A}'\mathbf{A}\}$. When the filter is stable and the input series $\boldsymbol{X}_t$ is stationary with cross-covariance matrices $\boldsymbol{\Gamma}_{xx}(l)$, the output $\boldsymbol{Z}_t = \sum_{j=-\infty}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{X}_{t-j}$ is a stationary process. The cross-covariance matrices of the stationary process $\{\boldsymbol{Z}_t\}$ are then given by

$$\boldsymbol{\Gamma}_{zz}(l) = \text{cov}[\boldsymbol{Z}_t, \boldsymbol{Z}_{t+l}] = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \boldsymbol{\Psi}_i \boldsymbol{\Gamma}_{xx}(l + i - j) \boldsymbol{\Psi}_j' \tag{A14.1.4}$$

It also follows, from (A14.1.1), that the spectral density matrix of the output $\boldsymbol{Z}_t$ has the representation

$$\mathbf{P}_{zz}(f) = \boldsymbol{\Psi}(e^{i2\pi f}) \mathbf{P}_{xx}(f) \boldsymbol{\Psi}(e^{-i2\pi f})' \tag{A14.1.5}$$

where $\mathbf{P}_{xx}(f)$ is the spectral density matrix of $\boldsymbol{X}_t$, and $\boldsymbol{\Psi}(z) = \sum_{j=-\infty}^{\infty} \boldsymbol{\Psi}_j z^j$ is the *transfer function* (matrix) of the linear filter. In addition, the cross-covariance matrices between $\boldsymbol{Z}_t$ and $\boldsymbol{X}_t$ are

$$\boldsymbol{\Gamma}_{zx}(l) = \text{cov}[\boldsymbol{Z}_t, \ \boldsymbol{X}_{t+l}] = \sum_{j=-\infty}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\Gamma}_{xx}(l + j)$$

and the cross-spectral density matrix between $\boldsymbol{Z}_t$ and $\boldsymbol{X}_t$ is

$$\mathbf{P}_{zx}(f) = 2 \sum_{l=-\infty}^{\infty} \boldsymbol{\Gamma}_{zx}(l) e^{-i2\pi f l} = \boldsymbol{\Psi}(e^{i2\pi f}) \mathbf{P}_{xx}(f)$$

so the transfer function $\boldsymbol{\Psi}(z)$ satisfies the relation $\boldsymbol{\Psi}(e^{i2\pi f}) = \mathbf{P}_{zx}(f)\mathbf{P}_{xx}(f)^{-1}$. In practice, when a causal linear filter is used to represent the relation between an observable input process $\boldsymbol{X}_t$ and an output process $\boldsymbol{Z}_t$ in a dynamic system, there will be added unobserved noise $\boldsymbol{N}_t$ in the system and a dynamic model of the form $\boldsymbol{Z}_t = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{X}_{t-j} + \boldsymbol{N}_t$ will be useful.

For a special example of the above linear filtering results, consider the basic stationary vector *white noise process* $\{\boldsymbol{a}_t\}$ defined in Section 14.1.3, with the properties that $E[\boldsymbol{a}_t] = 0$, $E[\boldsymbol{a}_t \boldsymbol{a}_t'] = \boldsymbol{\Sigma}$, and $E[\boldsymbol{a}_t \boldsymbol{a}_{t+l}'] = 0$ for $l \neq 0$. Hence, $\boldsymbol{a}_t$ has spectral density matrix $\mathbf{P}_{aa}(f) = 2\boldsymbol{\Sigma}$. Then the process $\boldsymbol{Z}_t = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t-j}$, with $\sum_{j=0}^{\infty} \|\boldsymbol{\Psi}_j\| < \infty$, is stationary and has cross-covariance matrices

$$\boldsymbol{\Gamma}_{zz}(l) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\Sigma} \boldsymbol{\Psi}_{j+l}' \tag{A14.1.6}$$

and spectral density matrix

$$\mathbf{P}_{zz}(f) = 2\boldsymbol{\Psi}(e^{i2\pi f})\boldsymbol{\Sigma}\boldsymbol{\Psi}(e^{-i2\pi f})' \tag{A14.1.7}$$

and the cross-covariance matrices between $\{\boldsymbol{Z}_t\}$ and $\{\boldsymbol{a}_t\}$ are $\boldsymbol{\Gamma}_{za}(l) = \boldsymbol{\Psi}_{-l}\boldsymbol{\Sigma}$ for $l \leq 0$ and zero for $l > 0$.

In addition, for the stationary VARMA(p, q) process with infinite MA representation (14.1.4), the covariance matrix-generating function is given by $\mathbf{G}(z) = \sum_{l=-\infty}^{\infty} \boldsymbol{\Gamma}(l) z^l = \boldsymbol{\Psi}(z^{-1})\boldsymbol{\Sigma}\boldsymbol{\Psi}(z)'$; hence, the spectral density matrix of the VARMA(p, q) process is given as in (A14.1.7) with $\boldsymbol{\Psi}(z) = \boldsymbol{\Phi}^{-1}(z)\boldsymbol{\Theta}(z)$.

EXERCISES

14.1. Consider the bivariate VMA(1) process $\boldsymbol{Z}_t = (\mathbf{I} - \boldsymbol{\Theta}B)\boldsymbol{a}_t$, with

$$\boldsymbol{\Theta} = \begin{bmatrix} 0.4 & 0.3 \\ -0.5 & 0.8 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$$

(a) Find the lag 0 and lag 1 autocorrelations and cross-correlations of $\boldsymbol{Z}_t$; that is, find the matrices $\boldsymbol{\Gamma}(0), \boldsymbol{\Gamma}(1)$, and $\boldsymbol{\rho}(0), \boldsymbol{\rho}(1)$.

(b) Find the individual univariate MA(1) models for z_{1t} and z_{2t}; that is, in the models $z_{it} = (1 - \eta_i B)\varepsilon_{it}, i = 1, 2$, find the values of the parameters η_i and $\sigma_{\varepsilon_i}^2 = \text{var}[\varepsilon_{it}]$, from $\boldsymbol{\Gamma}(0)$ and $\boldsymbol{\Gamma}(1)$.

(c) Show that the bivariate VMA(1) model above is invertible, state the matrix difference equation satisfied by the matrix weights $\boldsymbol{\Pi}_j$ in the infinite AR form of the VMA(1) model, and explicitly evaluate the $\boldsymbol{\Pi}_j$ for $j = 1, 2, 3, 4$.

(d) It follows from Section 14.5.2 that the diagonal elements of $\boldsymbol{\Sigma}$ represent the one-step-ahead forecast error variances for the two series when each series is forecast from the past history of both series, that is, when each series is forecast based on the bivariate model. Compare these one-step forecast error variances

in the bivariate model with the one-step forecast error variances $\sigma_{\varepsilon_i}^2$ based on the individual univariate models in (b).

14.2. For the stationary multivariate VAR(1) model $(\mathbf{I} - \mathbf{\Phi}B)\boldsymbol{Z}_t = \boldsymbol{a}_t$, it is known that $\mathbf{\Gamma}(0) - \mathbf{\Phi}\mathbf{\Gamma}(0)\mathbf{\Phi}' = \mathbf{\Sigma}$. Hence, if the model parameters $\mathbf{\Phi}$ and $\mathbf{\Sigma}$ are given, this matrix equation may be solved to determine $\mathbf{\Gamma}(0)$. In the bivariate case, this leads to three linear equations in the unknowns $\gamma_{11}(0), \gamma_{12}(0)$, and $\gamma_{22}(0)$. If these equations are expressed in matrix form as $\boldsymbol{A}[\gamma_{11}(0), \gamma_{12}(0), \gamma_{22}(0)]' = \boldsymbol{b}$, give explicitly the expressions for $\boldsymbol{A}$ and $\boldsymbol{b}$. Consider the specific case

$$\mathbf{\Phi} = \begin{bmatrix} 0.2 & 0.3 \\ -0.6 & 1.1 \end{bmatrix} \qquad \mathbf{\Sigma} = \begin{bmatrix} 4 & 1 \\ 1 & 1 \end{bmatrix}$$

(a) Show that

$$\mathbf{\Gamma}(0) = \begin{bmatrix} 5.667 & 4.000 \\ 4.000 & 10.667 \end{bmatrix}$$

Also, determine the stationarity of the VAR(1) model above, state the difference equation satisfied by the $\mathbf{\Gamma}(j), j \geq 1$, and find the values of $\mathbf{\Gamma}(1), \mathbf{\Gamma}(2)$, and $\mathbf{\Gamma}(3)$. In addition, compute the cross-correlation matrices $\boldsymbol{\rho}(0), \boldsymbol{\rho}(1)$, $\boldsymbol{\rho}(2)$, and $\boldsymbol{\rho}(3)$.

(b) Find the matrix coefficients $\mathbf{\Psi_1}, \mathbf{\Psi_2}$, and $\mathbf{\Psi_3}$ in the infinite MA representation for $\boldsymbol{Z}_t$, and hence, compute the covariance matrix of the bivariate lead l forecast errors from the bivariate model using the formula $\mathbf{\Sigma}(l) = \sum_{j=0}^{l-1} \mathbf{\Psi}_j \mathbf{\Sigma} \mathbf{\Psi}_j'$, for $l = 1, 2, 3$.

(c) For a bivariate VAR(1) model, indicate what simplifications occur in the model when $\mathbf{\Phi}$ is lower triangular (i.e., $\phi_{12} = 0$). In particular, show in this case that the bivariate system can be expressed equivalently in the form of a ''unidirectional'' transfer function model, as in Chapter 12, with z_{1t} as input series and z_{2t} as output. In addition, indicate the specific nature of the univariate ARMA model for the series z_{2t} implied by this situation.

(d) For a bivariate VAR(1) model, show that the case $\det(\mathbf{\Phi}) = 0$ implies that there exists a linear combination of z_{1t} and $z_{2t}, Y_{1t} = c_{11}z_{1t} + c_{12}z_{2t}$, which is a white noise series, and a second linear combination $Y_{2t} = c_{21}z_{1t} + c_{22}z_{2t}$, which is again a univariate AR(1) process.

Hint: If $\det(\mathbf{\Phi}) = 0$, then $\mathbf{\Phi}$ has rank at most one and can be written as

$$\mathbf{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \alpha\phi_{11} & \alpha\phi_{12} \end{bmatrix} = \begin{bmatrix} 1 \\ \alpha \end{bmatrix} \begin{bmatrix} \phi_{11} & \phi_{12} \end{bmatrix}$$

14.3. Consider the VAR(p) model

$$\boldsymbol{Z}_t = \mathbf{\Phi}_1 \boldsymbol{Z}_{t-1} + \mathbf{\Phi}_2 \boldsymbol{Z}_{t-2} + \cdots + \mathbf{\Phi}_p \boldsymbol{Z}_{t-p} + \boldsymbol{a}_t$$

Verify that the model can be expressed as a VAR(1) model in terms of the kp-dimensional vector $\boldsymbol{Y}_t = (\boldsymbol{Z}_t', \boldsymbol{Z}_{t-1}', \ldots, \boldsymbol{Z}_{t-p+1}')'$, $\boldsymbol{Y}_t = \boldsymbol{\Phi}\boldsymbol{Y}_{t-1} + \boldsymbol{e}_t$, using the $kp \times kp$ *companion matrix* $\boldsymbol{\Phi}$ for the VAR(p) operator $\boldsymbol{\Phi}(B) = \boldsymbol{I} - \boldsymbol{\Phi}_1 B - \cdots - \boldsymbol{\Phi}_p B^p$,

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_2 & \cdots & \cdots & \boldsymbol{\Phi}_p \\ \mathbf{I} & 0 & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \cdot \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix}$$

In addition, show that $\det\{\mathbf{I} - \boldsymbol{\Phi}B\} = \det\{\mathbf{I} - \boldsymbol{\Phi}_1 B - \cdots - \boldsymbol{\Phi}_p B^p\}$, and hence the stationarity condition for the VAR(p) process is equivalent to the condition that all eigenvalues of the companion matrix $\boldsymbol{\Phi}$ be less than one in absolute value. (*Hint*: To evaluate $\det\{\mathbf{I} - \boldsymbol{\Phi}B\}$, multiply the ith column of $\mathbf{I} - \boldsymbol{\Phi}B$ by B and add to the $(i-1)$st column, successively, for $i = p, p-1, \ldots, 2$.)

14.4. For a bivariate VAR(2) model $\boldsymbol{Z}_t = \boldsymbol{\Phi}_1 \boldsymbol{Z}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{Z}_{t-2} + \boldsymbol{a}_t$, with

$$\boldsymbol{\Phi}_1 = \begin{bmatrix} 1.5 & -0.6 \\ 0.3 & 0.2 \end{bmatrix} \qquad \boldsymbol{\Phi}_2 = \begin{bmatrix} -0.5 & 0.3 \\ 0.7 & -0.2 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$$

(a) Verify that this model is stationary based on the nature of the roots of $\det\{\mathbf{I} - \boldsymbol{\Phi}_1 B - \boldsymbol{\Phi}_2 B^2\} = 0$. (Note that you may want to make use of the result of Exercise 14.3 for computational convenience.)

(b) Calculate forecasts $\hat{\boldsymbol{Z}}_n(l)$ for $l = 1, \ldots, 5$ steps ahead, given that $\boldsymbol{Z}_n = (1.2, 0.6)'$ and $\boldsymbol{Z}_{n-1} = (0.5, 0.9)'$.

(c) Find the coefficient matrices $\boldsymbol{\Psi}_j$, $j = 1, \ldots, 4$, in the infinite MA representation of the process, and find the forecast error covariance matrices $\boldsymbol{\Sigma}(l)$ for $l = 1, \ldots, 5$.

14.5. Consider the simple transfer function model

$$(1 - B)z_{1t} = \varepsilon_{1t} - \theta\varepsilon_{1,t-1} \qquad z_{2t} = \omega z_{1t} + \varepsilon_{2t}$$

where ε_{1t} and ε_{2t} are independent white noise processes.

(a) Determine the univariate ARIMA model for z_{2t}, and note that z_{2t} is nonstationary.

(b) Express the bivariate model for $\boldsymbol{Z}_t = (z_{1t}, z_{2t})'$ in the general form of a ''generalized'' ARMA(1, 1) model, $(\mathbf{I} - \boldsymbol{\Phi}_1 B)\boldsymbol{Z}_t = (\mathbf{I} - \boldsymbol{\Theta}_1 B)\boldsymbol{a}_t$, and determine that one of the eigenvalues of $\boldsymbol{\Phi}_1$ is equal to one.

(c) Determine the bivariate model for the first differences $(1 - B)\boldsymbol{Z}_t$, and show that it has the form of a bivariate IMA(1, 1) model, $(1 - B)\boldsymbol{Z}_t = (\mathbf{I} - \boldsymbol{\Theta}^* B)\boldsymbol{a}_t$,

where the MA operator $(\mathbf{I} - \boldsymbol{\Theta}^* B)$ is not invertible. Hence, this model represents an ''overdifferencing'' of the bivariate series $\boldsymbol{Z}_t$.

14.6. Suppose $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$, with $N = 60$, is a sample from a bivariate VAR(1) process, with sample covariance matrices obtained as

$$\hat{\boldsymbol{\Gamma}}(0) = \begin{bmatrix} 1.0 & 1.0 \\ 1.0 & 2.0 \end{bmatrix} \qquad \hat{\boldsymbol{\Gamma}}(1) = \begin{bmatrix} 0.6 & 0.4 \\ 0.7 & 1.2 \end{bmatrix} \qquad \hat{\boldsymbol{\Gamma}}(2) = \begin{bmatrix} 0.30 & 0.10 \\ 0.42 & 0.64 \end{bmatrix}$$

(a) Obtain the corresponding estimated correlation matrices $\hat{\boldsymbol{\rho}}(0), \hat{\boldsymbol{\rho}}(1)$, and $\hat{\boldsymbol{\rho}}(2)$.

(b) Find the sample Yule–Walker estimates for $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ in the VAR(1) model, and find an estimate for the approximate covariance matrix of the estimator $\hat{\boldsymbol{\Phi}}$, that is, for the covariance matrix of $\text{vec}[\hat{\boldsymbol{\Phi}}']$.

(c) Based on the results in (b), test whether the matrix $\boldsymbol{\Phi}$ has a lower triangular structure; that is, test whether $\phi_{12} = 0$.

14.7. Suppose that a three-dimensional VARMA process $\boldsymbol{Z}_t$ has Kronecker indices $K_1 = 3, K_2 = 1$, and $K_3 = 2$.

(a) Write the form of the coefficient matrices $\boldsymbol{\Phi}_j^{\#}$ and $\boldsymbol{\Theta}_j^{\#}$ in the echelon canonical ARMA model structure of equation (14.7.1) for this process.

(b) For this process $\{\boldsymbol{Z}_t\}$, describe the nature of the zero canonical correlations that occur in the canonical correlation analysis of the past vector $\boldsymbol{P}_t = (\boldsymbol{Z}_t', \boldsymbol{Z}_{t-1}', \ldots)'$ and various future vectors $\boldsymbol{F}_{t+1}^*$.

(c) Write the form of the minimal dimension echelon state-space model corresponding to the echelon canonical ARMA model for this process.

14.8. Verify that any VAR(p) model $\boldsymbol{Z}_t = \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{Z}_{t-j} + \boldsymbol{a}_t$ can be expressed equivalently in the error correction form of equation (14.8.2) as $\boldsymbol{W}_t = \boldsymbol{C}\boldsymbol{Z}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Phi}_j^* \boldsymbol{W}_{t-j} + a_t$, where $\boldsymbol{W}_t = \boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}, \boldsymbol{\Phi}_j^* = -\sum_{i=j+1}^{p} \boldsymbol{\Phi}_i$, and $\boldsymbol{C} = -\boldsymbol{\Phi}(1) = -\left(\boldsymbol{I} - \sum_{j=1}^{p} \boldsymbol{\Phi}_j\right)$.

14.9. Express the model for the nonstationary bivariate process $\boldsymbol{Z}_t$ given in Exercise 14.5 in an error correction form, similar to equation (14.8.2), as $\boldsymbol{W}_t = \mathbf{C}\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t - \boldsymbol{\Theta}a_{t-1}$, where $\boldsymbol{W}_t = (1 - B)\boldsymbol{Z}_t$. Determine the structure (and the ranks) of the matrices $\boldsymbol{C}$ and $\boldsymbol{\Theta}$ explicitly.

14.10. Consider analysis of the logarithms of monthly flour price indices from three U.S. cities. The raw (unlogged) data were given and analyzed by Tiao and Tsay (1989). To first investigate a VAR model for these data, with possible reduced-rank structure, the results of the partial canonical correlation analysis of Section 14.7.3, in terms of the (squared) partial canonical correlations $\hat{\rho}_i^2(j)$ between $\boldsymbol{Z}_t$ and $\boldsymbol{Z}_{t-j}$ for lags $j = 1, \ldots, 6$, and the associated test statistic values computed using (14.7.8) are displayed in the following table:

j	Squared Correlations			$C(j, r)$ $r = 2$	$r = 1$	$r = 0$
1	0.747,	0.814,	0.938	129.93	288.89	551.67
2	0.003,	0.081,	0.274	0.29	7.97	36.91
3	0.001,	0.007,	0.035	0.07	0.69	3.76
4	0.000,	0.015,	0.047	0.03	1.29	5.31
5	0.017,	0.036,	0.073	1.36	4.24	10.22
6	0.000,	0.020,	0.077	0.00	1.51	7.49

In addition, values of $|\tilde{\mathbf{\Sigma}}_j|$ and of the AIC_j and HQ_j model selection criteria for the full-rank VAR(j) models are given as follows:

j (AR order)	1	2	3	4	5	6
$\lvert\tilde{\mathbf{\Sigma}}_j\rvert(\times 10^{-10})$	1.66213	1.12396	1.10523	1.06784	0.88963	0.81310
AIC_j	−22.336	−22.542	−22.369	−22.210	−22.195	−22.084
HQ_j	−22.240	−22.350	−22.079	−21.822	−21.707	−21.494

Interpret the preliminary model specification information above, and specify the structure (order and possible reduced ranks) of a VAR model that may seem appropriate for these data based on these results.

PART FOUR

DESIGN OF DISCRETE CONTROL SCHEMES

In earlier chapters we studied the modeling of discrete univariate time series and dynamic systems involving two or more time series. We saw how once adequate models have been developed, they can be used to generate forecasts of future observations, to characterize the transfer function of a dynamic system, and to represent the interrelationships among several time series of a multivariate dynamic system. Examples involving real-world applications have been used for illustration. However, the models and the methodology are of much wider importance than even these applications indicate. The ideas we have outlined are of importance in the analysis of a wide class of stochastic–dynamic systems occurring, for example, in economics, engineering, commerce, hydrology, meteorology, and in organizational studies.

It is obviously impossible to illustrate every application. Rather, it is hoped that the theory and examples of this book will help the reader to adapt the general methodology to their own particular problems. In doing this, the dynamic and stochastic models we have discussed will often act as *building blocks* that can be linked together to represent the particular system under study. The techniques of identification, estimation, and diagnostic checking, similar to those we have illustrated, should be useful to establish the model. Finally, recursive calculations and the ideas considered under the general heading of forecasting will have wider application in evaluating the adequacy and the usefulness of a model for a specific purpose once the model has been fitted.

We shall conclude this book by illustrating these possibilities in one further application—the design of feedback and feedforward control schemes. In working through Chapter 15, it is the task of bringing together the previously discussed ideas in a fresh application, quite as much as the detailed results, that we hope will be of value.

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

15

ASPECTS OF PROCESS CONTROL

The term *process control* is used in different ways. Shewhart charts and other quality control charts are frequently employed in industries concerned with the manufacture of discrete ''parts'' in what is called *statistical process control* (SPC). By contrast, various forms of feedback and feedforward adjustment are used, particularly in the process and chemical industries, in what we call *engineering process control* (EPC). Because the adjustments made by engineering process control are usually computed and applied automatically, this type of control is sometimes called *automatic process control* (APC). However, the manner in which adjustments are applied is a matter of convenience, so we will not use that terminology here. The object of this chapter is to draw on the earlier discussions in this book to provide insight into the statistical aspects of these control methods and to appreciate better their relationships and objectives.

We first discuss *process monitoring* using, for example, Shewhart control charts and contrast this with techniques for *process adjustment*. In particular, a common adjustment problem is to maintain an output variable close to a target value in a dynamic system subject to disturbances by manipulation of an input variable, to obtain feedback control. *Feedback control* schemes use only the observed deviation of the output from target as a basis for adjustment of the input variable. We consider this problem first in a purely intuitive way and then relate this to some of the previously discussed stochastic and transfer function models to yield feedback control schemes producing minimum mean square error (MMSE) at the output. This leads to a discussion of discrete schemes, which are analogs of the proportional–integral (PI) schemes of engineering control, and we show how simple charts may be devised for *manually adjusting* processes with PI control.

It turns out that minimum mean square error control often requires excessively large adjustments of the input variable. ''Optimal'' constrained schemes are, therefore, introduced that require much smaller adjustments at the expense of only minor increases in

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

the output mean square error. These constrained schemes are generally not PI schemes, but in certain important cases it turns out that appropriately chosen PI schemes can often closely approximate their behavior. Particularly, in industries concerned with the manufacture of parts, there may be a fixed cost associated with adjusting the process and, in some cases, a monitoring cost associated with obtaining an observation. We therefore also discuss bounded adjustment schemes for feedback control that minimize overall cost in these circumstances.

In some instances, one or more *sources* of disturbance may be measured, and these measurements may be used to compensate potential deviations in the output. This type of adjustment action is called *feedforward control*, as compared to feedback control where only the observed deviation of output from target is used as a basis for adjustment. In certain instances it may also be desirable to use a combination of these two modes of control, and this is referred to as *feedforward–feedback control*. We therefore also present feedforward and feedforward–feedback types of control schemes for a general dynamic system that yield minimum mean square error at the output. Finally, we consider a general procedure for monitoring control schemes for possible changes in parameter values using Cuscore charts. More general discussion is given in the appendices and references.

15.1 PROCESS MONITORING AND PROCESS ADJUSTMENT

Process control is no less than an attempt to cancel out the effect of a fundamental physical law—the second law of thermodynamics, which implies that if left to itself, the entropy or disorganization of any system can never decrease and will usually increase. SPC and EPC are two complementary approaches to combat this law. SPC attempts to *remove* disturbances using *process monitoring*, while EPC attempts to *compensate* them using *process adjustment* (see also Box and Kramer, 1992).

15.1.1 Process Monitoring

The SPC strategy for stabilization of a process is to standardize procedures and raw materials and to use hypothesis-generating devices (such as graphs, check sheets, Pareto charts, cause–effect diagrams, etc.) to track down and eliminate causes of trouble (see, for example, Ishikawa, 1976). Since searching for assignable causes is tedious and expensive, it usually makes sense to wait until ''statistically significant'' deviations from the stable model occur before instituting this search. This is achieved by the use of process *monitoring charts* such as Shewhart charts, Cusum charts, and Roberts' EWMA charts. The philosophy is ''don't fix it when it ain't broke''—don't needlessly tamper with the process (see, for example, Deming, 1986).

Figure 15.1 shows an example of process monitoring using a Shewhart control chart. Condoms were tested by taking a sample of 50 items every 2 hours from routine production, inflating them to a very high fixed pressure, and noting the proportion that burst. Figure 15.1 shows data taken during the startup of a machine making these articles. Studies from similar machines had shown that a high-quality product was produced if the proportion failing this very severe test was $p = 0.20$.

The *reference* distribution indicated by the bars on the right of Figure 15.1(a) characterizes desired process behavior. It is a binomial distribution showing the probabilities of getting various proportions failing in random samples of $n = 50$ when p stays constant at a

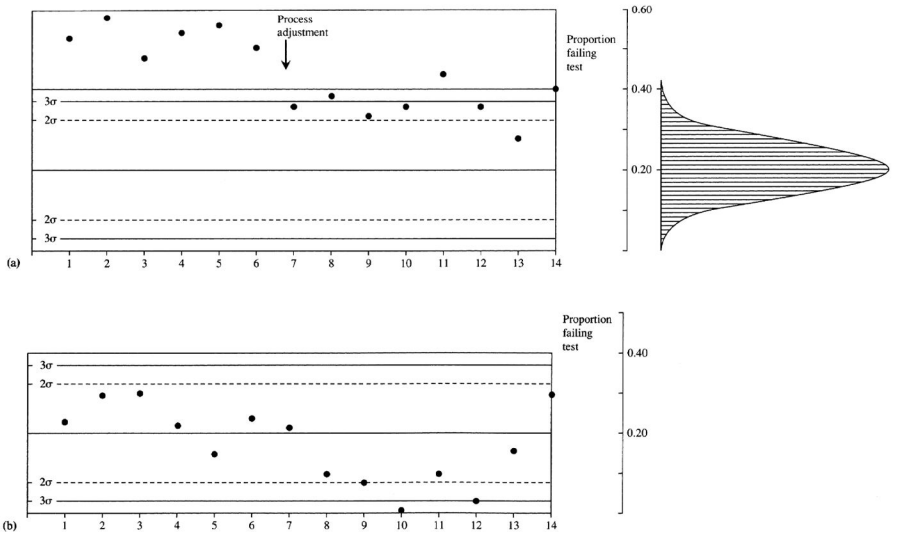

FIGURE 15.1 Shewhart charts for the proportion of condoms failing an inflation test. (a) Data taken before the process was brought to a state of control. (b) Data taken at a later stage of development.

value of 0.20. If the data behaved like a random sequence from this reference distribution, we should say the process appeared to be in a *state of control* and no action would be called for. By contrast, if the data did not have this appearance, showing outliers or suspicious patterns, we might have reason to suppose that something else was going on. In practice, the whole reference distribution would not usually be shown. Instead, upper and lower control limits and warning lines would be set. When, as in this case, a normal curve (shown as a continuous line) provides a close approximation to the reference distribution, these are usually set at $\pm 2\sigma$ and $\pm 3\sigma$ with $\sigma = \sqrt{p(1-p)/n}$, the standard deviation of the sample proportion from a binomial distribution. In this example, with $p = 0.20$ and $n = 50$, this gives $\sigma = 0.057$. Figure 15.1(a) shows that during the startup phase, the process was badly out of control, with the proportion of items failing the test initially as high as 50%. A process adjustment made after 12 hours of operation brought the proportion of defectives down to around 40%, but further changes were needed to get the process to a state of control at the desired level of $p = 0.20$. By a series of management actions, this was eventually achieved and Figure 15.1(b) shows the operation of the process at a later stage of development. Although for the most part the system now appears to be in the desired state of control, notice that the 10th point on the chart fell *below the lower* $\pm 3\sigma$ line. Subsequent investigation showed that the testing procedure was responsible for this aberrant point. A fault in the air line had developed and the condoms tested at about this time were inadvertently submitted to a much reduced air pressure, resulting in a falsely low value of the proportion defective. Corrective action was taken and the system was modified so that the testing machine would not function unless the air pressure was at the correct setting, ensuring that this particular fault could not occur again.

Monitoring procedures of this kind are obviously of great value. Following Shewhart (1931) and Deming (1986), we refer to the natural variation in the process when in state of control (binomial variation for a sample of $n = 50$ with $p = 0.20$ in this case) as due to *common causes*. The common cause system can only be changed by management action that alters the system. Thus, a new type of testing machine might be introduced for which the acceptable proportion of defects should be 10%. Common cause variation would then be binomial about the value $p = 0.10$.

The fault in the air line that was discovered by using the chart is called a *special*[1] *cause*. By suitable ''detective'' work, it is often possible for the plant operators to track down and eliminate special causes. The objectives of process monitoring are thus (1) to establish and continually confirm that the desired common cause system remains in operation and (2) to look for deviations unlikely to be due to chance that can lead to the tracking and elimination of assignable causes of trouble.

15.1.2 Process Adjustment

Although we must always make a dedicated endeavor to remove causes of variation such as unsatisfactory testing methods, differences in raw materials, differences in operators, and so on, some processes cannot be fully brought to a satisfactory state of stability in this way. Despite our best efforts, there remains a tendency for the process to wander off

[1] Also called an ''assignable'' cause. However, we are sometimes faced with a system that is demonstrably not in a state of control and yet no causative reason can be found. So we will stay with Deming in his less optimistic word ''special.''

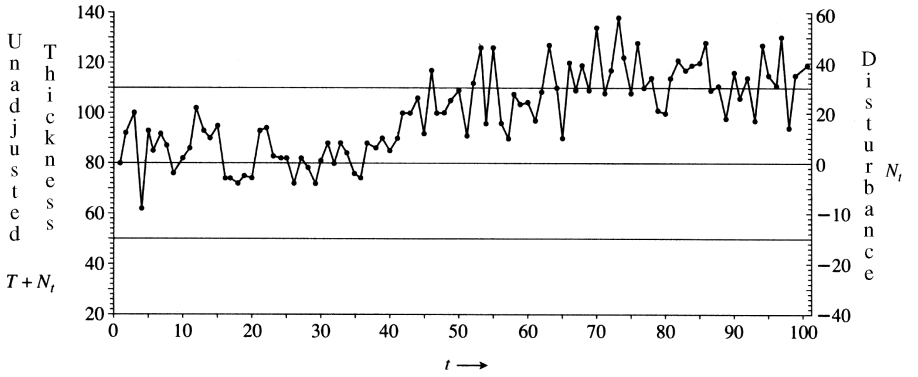

FIGURE 15.2 One hundred successive values of the thickness of a metallic film when no adjustment was applied.

target. This may be due to known but uncontrollable phenomena such as variations in ambient temperature, humidity, and feedstock quality, or due to causes currently unknown. In such circumstances, some system of process *adjustment* or *regulation* may be necessary in which manipulation of some additional variable is used to *compensate* for deviations in the quality characteristic.

To fix ideas, we first introduce a simple feedback adjustment scheme relying on a purely empirical argument and leave theoretical justification until later. Consider the measurements shown in Figure 15.2 of the thickness of a very thin metallic film taken at equally spaced units of time. The quality characteristic was badly out of control, but standard procedures failed to stabilize it (Box, 1991a). Suppose that the *disturbance* N_t is defined as the deviation of this quality characteristic from its target value T when *no adjustment is made*; that is, N_t is the underlying noise process. Suppose also that there is a manipulable variable—deposition rate X—which can be used conveniently to adjust the thickness, and that a unit change in X will produce g units of change in thickness and will take full effect in one time interval. If at time t, X was set equal to X_t, then at time $t + 1$ the deviation from target, $\varepsilon_{t+1} = Y_{t+1} - T$, after adjustment would be

$$\varepsilon_{t+1} = gX_t + N_{t+1} \tag{15.1.1}$$

Now suppose that *at time t* you can, in some way or the other, compute an estimate (forecast) $\hat{N}_t(1)$ of N_{t+1} and that this forecast has an error $e_t(1)$, so that

$$N_{t+1} = \hat{N}_t(1) + e_t(1) \tag{15.1.2}$$

Then using (15.1.1) and (15.1.2),

$$\varepsilon_{t+1} = gX_t + \hat{N}_t(1) + e_t(1) \tag{15.1.3}$$

If, in particular, X can be adjusted so that at time t,

$$X_t = -\frac{1}{g}\hat{N}_t(1) \tag{15.1.4}$$

then for the adjusted process

$$\varepsilon_{t+1} = e_t(1) \tag{15.1.5}$$

Thus, the deviation from target ε_{t+1} for the *adjusted* process would now be the error $e_t(1)$ in *forecasting* N_{t+1}, instead of the deviation N_{t+1} measured when the process is not adjusted.

If we used measurements of one or more of the known disturbing *input* factors (e.g., ambient temperature) to calculate the estimate $\hat{N}_t(1)$ of N_{t+1}, we would have an example of feedforward control. If the estimate $\hat{N}_t(1)$ of N_{t+1} directly or indirectly used only present and past values of the *output* disturbance $N_t, N_{t-1}, N_{t-2}, \ldots$, equation (15.1.4) would define a system of *feedback* control. A system of mixed *feedback–feedforward* control would employ both kinds of data. For simplicity, we will focus on the feedback case in the next three sections, and consider feedforward and mixed control in Section 15.5.

15.2 PROCESS ADJUSTMENT USING FEEDBACK CONTROL

Empirical Introduction. It might often be reasonable to use for the estimate $\hat{N}_t(\mathrm{l})$ in (15.1.4) some kind of weighted average of past values $N_t, N_{t-1}, N_{t-2}, \ldots$. In particular, an *exponentially* weighted moving average (EWMA) has intuitive appeal since recently occurring data are given most weight. Suppose, then, that $\hat{N}_t(\mathrm{l})$ is an EWMA,

$$\hat{N}_t(\mathrm{l}) = \lambda(N_t + \theta N_{t-1} + \theta^2 N_{t-2} + \cdots) \qquad 0 \le \theta \le 1 \tag{15.2.1}$$

where θ is the smoothing constant and $\lambda = 1 - \theta$,

We first consider the situation where, as has usually been the case in the process industries, adjustments are continually made as each observation comes to hand. Then using equation (15.1.4), the *adjustment* (*change* in deposition rate) made at time t would be given by

$$X_t - X_{t-l} = -\frac{1}{g}[\hat{N}_t(\mathrm{l}) - \hat{N}_{t-1}(1)] \tag{15.2.2}$$

Now with $e_{t-1}(1) = N_t - \hat{N}_{t-1}(1)$ the forecast error, the updating formula for an EWMA forecast can be written as

$$\hat{N}_t(1) - \hat{N}_{t-1}(1) = \lambda e_{t-1}(1) \tag{15.2.3}$$

Therefore, for any feedback scheme in which the compensatory variable X was set so as to cancel out an EWMA of the noise $\{N_t\}$, the required adjustment should be such that

$$X_t - X_{t-1} = -\frac{\lambda}{g} e_{t-1}(1) = -\frac{\lambda}{g}\varepsilon_t \tag{15.2.4}$$

For the metal deposition process, $g = 1.2$, $\lambda = 0.2$, and $T = 80$, so that the adjustment equation is

$$X_t - X_{t-1} = -\frac{1}{6}\varepsilon_t \tag{15.2.5}$$

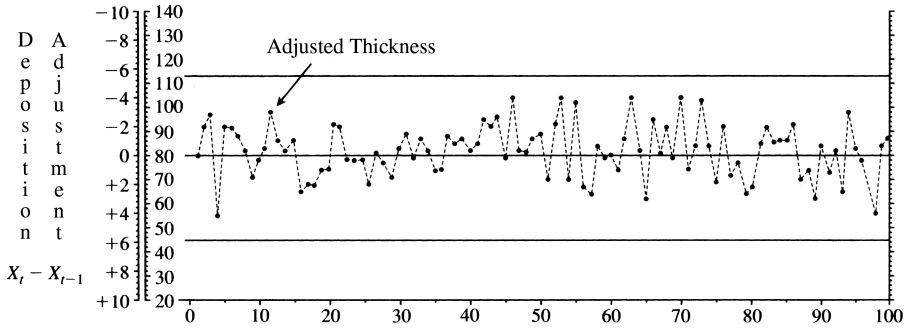

FIGURE 15.3 Manual adjustment chart for thickness that allows the operator to read off the appropriate change in deposition rate.

15.2.1 Feedback Adjustment Chart

This kind of adjustment is very easily applied, as is shown in Figure 15.3. This shows a manual feedback adjustment chart (Box and Jenkins, 1976) for the metallic thickness example given previously. To use it, the operator records the latest value of thickness and reads off on the adjustment scale the appropriate amount by which he or she should now increase or decrease the deposition rate. For example, the first recorded thickness of 80 is on target, so no action is called for. The second value of 92 is 12 units above the target, so $\varepsilon_2 = 12$, corresponding on the left-hand scale to a deposition rate adjustment of $X_2 - X_1 = -2$. Thus, the operator should now reduce the deposition rate by 2 units from its previous level.

Notice that the successive recorded thickness values shown on this chart are the readings that would actually occur *after adjustment*; the underlying disturbance is, of course, not seen on this chart. In this example, over the recorded period of observation, the chart produces a more than fivefold reduction in mean square error; the standard deviation of the adjusted thickness being now only about $\sigma_\varepsilon = 11$. Notice the following:

1. The chart is no more difficult to use than a Shewhart chart.
2. While the ''intuitive'' adjustment would be $-(1/g)\varepsilon_t = -(5/6)\varepsilon_t$ (corresponding to what Deming called ''tinkering''), the adjustment given by equation (15.2.4) is $-(\lambda/g)\varepsilon_t = -(1/6)\varepsilon_t$. Thus, it uses a discounted or ''damped'' estimate $\lambda\varepsilon_t$ of the deviation from target to determine the appropriate adjustment, where the discount factor λ is $1 - \theta$, with θ being the smoothing constant of the EWMA estimate of the noise.
3. By summing equation (15.2.4), we see that the *total* adjustment at time t is

$$X_t = k_0 + k_I \sum_{i=1}^{t} \varepsilon_i \tag{15.2.6}$$

with $k_0 = X_0$ and $k_I = -\lambda/g$. This adjustment procedure thus depends on the *cumulative* sum of the adjustment errors ε_i and the constant k_I determines how much the ''intuitive'' adjustment is discounted.

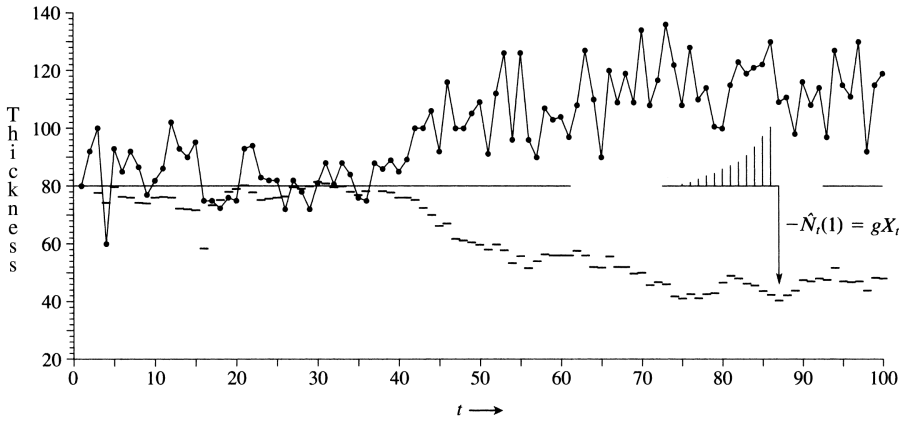

FIGURE 15.4 Dashes indicate the total adjustment $\hat{N}_t(1) = -gX_t$ achieved by the manual adjustment chart of Figure 15.3.

4. It follows from the previous argument that the adjustment is also equivalent to estimating at each time t the next value of the total unadjusted disturbance N_{t+1} by an *exponentially weighted average* of its past values and using this estimate to make an appropriate adjustment. This is illustrated for the metallic thickness example in Figure 15.4. Notice that in this preliminary discussion we have not explicitly assumed any particular time series model or claimed any particular optimal properties for the procedure. That the procedure can be discussed in such terms accounts, to some extent, for its remarkable robustness, which we discuss later.

In summary, then:

1. By *process monitoring* we mean the use of, for example, Shewhart charts and/or Cusum or Cuscore charts, as discussed by Box and Ramírez (1992). These are devices for continually checking a model that represents the desired ideal stable state of the system: for example, normal, independent, identically distributed (iid) variation about a fixed target T. The use of such charts can lead to the elimination of special causes pointed to by discrepant behavior. The judgment that behavior is sufficiently discrepant to merit attention is decided by a process analogous to *hypothesis testing*. Its properties are described in terms of probabilities (e.g., the probability of a point falling outside the 3σ limits of a Shewhart chart).
2. By *process adjustment* we mean the use of feedback and feedforward control or some combination of these to maintain the process as close as possible to some desired target value. Process adjustment employs a system of statistical *estimation* (forecasting) rather than of hypothesis testing, and its properties are described, for example, by output *mean square error*. Process monitoring and process adjustment are complementary rather than competitive corresponding to the complementary roles of hypothesis testing and estimation (see, for example, Box, 1980). We discuss this point more fully later in the chapter.

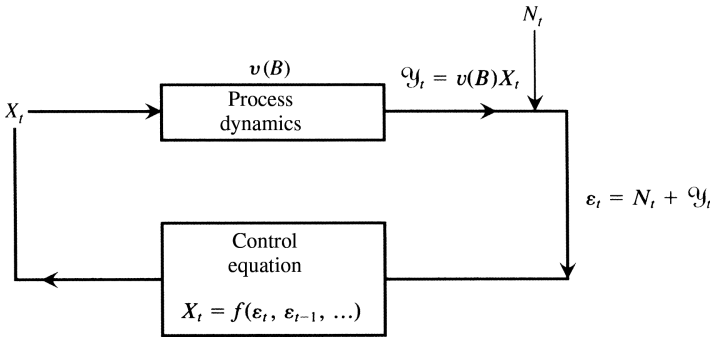

FIGURE 15.5 Feedback control loop.

15.2.2 Modeling the Feedback Loop

A somewhat more general system of feedback control is shown in Figure 15.5. The process is affected by a disturbance that in the absence of compensatory action would cause the output quality characteristic to deviate from target by an amount N_t. Thus, $\{N_t\}$ is a time series exemplifying what would happen at the output if no control were applied. In fact, a compensating variable X_t (deposition rate in our example) can be manipulated to cancel out this disturbance as far as possible. Changes in X will pass through the process and be acted on by its dynamics to produce at time t an amount or compensation $\mathcal{Y}_t$ at the output (again measured as a deviation from target). To the extent that this compensation $\mathcal{Y}_t$ fails to cancel out the disturbance N_t, there will be an error, or deviation from target $\varepsilon_t = Y_t - T$, equal to $\varepsilon_t = N_t + \mathcal{Y}_t$. The controller is some means (automatic or manual) that brings into effect the control equation $X_t = f(\varepsilon_t, \varepsilon_{t-1}, \ldots)$, which adjusts the output depending on present and past errors.

A device that has been used in the process industries for many years is the *three-term controller*. Controllers of this kind are usually operated automatically and employ continuous rather than discrete measurement and adjustment. If ε_t is the error at the output at time t, control action could, in particular, be made proportional to ε itself, to its integral with respect to time, or to its derivative with respect to time. A three-term controller uses a linear combination of these modes of control action, so that if X_t indicates the level of the manipulated variable at time t, the control equation is of the form

$$X_t = k_0 + k_D \frac{d\varepsilon_t}{dt} + k_P \varepsilon_t + k_I \int \varepsilon_t \, dt \tag{15.2.7}$$

where k_D, k_P, and k_I are constants.

Frequently, only one or two of these three modes of action are used. In particular, if only k_I is nonzero ($k_D = 0, k_P = 0$), we have *integral* control. If only k_I and k_P are nonzero ($k_D = 0$), we have *proportional--integral* (PI) control.

Notice that in the example we have just discussed, where the result of any adjustment fully takes effect at the output in one time interval, the dynamics of the process are represented by $\mathcal{Y}_t = gX_{t-1} = gBX_t$. The control equation $X_t = k_0 + k_I \sum_{i=1}^{t} \varepsilon_i$ in (15.2.6) is then the discrete analog of the control engineer's *integral* control.

In general, the discrete analog of (15.2.7) is

$$X_t = k_0 + k_D \nabla \varepsilon_t + k_P \varepsilon_t + k_I \sum_{i=1}^{t} \varepsilon_i$$

or in terms of the adjustment to be made,

$$\begin{aligned} x_t = X_t - X_{t-1} &= k_D \nabla^2 \varepsilon_t + k_P \nabla \varepsilon_t + k_I \varepsilon_t \\ &= c_1 \varepsilon_t + c_2 \varepsilon_{t-1} + c_3 \varepsilon_{t-2} \end{aligned}$$

where c_1, c_2, and c_3 are suitable constants. Not unexpectedly, control equations of this type are of considerable practical value.

15.2.3 Simple Models for Disturbances and Dynamics

So far we introduced a simple system of feedback control on purely empirical grounds. The *efficiency* of any such system will depend on the nature of the disturbance and the dynamics of the process. From a theoretical point of view, we can consider very general models for noise and dynamics and then proceed to find the control equation that ''optimizes'' the system in accordance with some criterion. However, the practical effectiveness of such models is usually determined by whether they, and the ''optimization'' criterion, make broad *scientific* sense and by their robustness to likely deviations from the ideal. We have already kept this in mind when discussing control procedures from a purely commonsense point of view and we will continue to do so when choosing models for the disturbance and for process dynamics.

Characterizing Appropriate Disturbance Models with a Variogram. A tool that helps to characterize process disturbances is the standardized variogram, which measures the variance of the difference between observations m steps apart compared to the variance of the difference of observations one step apart:

$$G_m = \frac{\text{var}[N_{t+m} - N_t]}{\text{var}[N_{t+1} - N_t]} \equiv \frac{V_m}{V_1} \tag{15.2.8}$$

For a stationary process, G_m is a simple function of the autocorrelation function. In fact, then, $G_m = (1 - \rho_m)/(1 - \rho_1)$. However, the variogram can be used to characterize nonstationary as well as stationary behavior. Figure 15.6 shows realizations of 100 observations initially on target generated by (a) a white noise process, (b) a first-order autoregressive process, and (c)–(f) IMA(0, 1, 1) processes with $\lambda = 0.1, 0.2, 0.3, 0.4$, respectively. The corresponding theoretical standardized variograms for these time series models are also shown.

In some imaginary world we might, once and for all, set the controls of a machine and give a set of instructions to an ever-alert and never-forgetting operator, and this would yield a perfectly stable process from that point on. In such a case the disturbance might be represented by a ''white noise'' series, and its corresponding standardized variogram G_m would be independent of m and equal to 1. But, in reality, left to themselves, machines involved in production are slowly losing adjustment and wearing out, and left to themselves, people tend, gradually, to forget instructions and miscommunicate. Thus, for an *uncontrolled* disturbance, some kind of monotonically increasing variogram would be expected. We cannot obtain such a variogram from a linear *stationary model*, for although

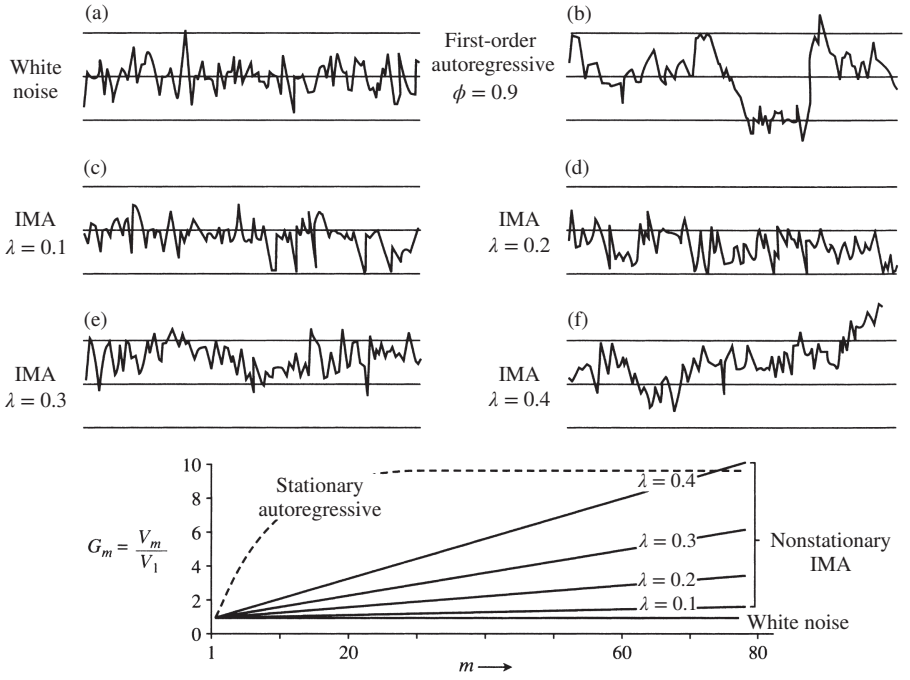

FIGURE 15.6 Realization of white noise, autoregressive, and IMA(0, 1, 1) time series with theoretical variograms.

G_m can initially increase with m, it will always approach an asymptote for such a process. That this can happen quite quickly, even when successive observations are highly positively correlated, is illustrated by the variogram shown in the figure for the first-order stationary autoregressive time series model $N_t = 0.9N_{t-1} + a_t$. In this example, even though successive deviations N_t from the target value have autocorrelation 0.9, G_m is already within 5% of its asymptotic value after only 20 lags. This implies that, for example, when generated by such a model, observations 100 steps apart differ little more than those 20 steps apart.

A model that can approximate the behavior of an uncontrolled system that *continuously* increases its entropy may be arrived at by thinking of the disturbance as containing two parts, a transitory part b_t and a nontransitory part z_t:

$$N_t = b_t + z_t \tag{15.2.9}$$

The transitory part b_t is associated only with the tth observation and is supposed independent of observations taken at every other time. Typical sources contributing to b_t are measurement and sampling errors. We represent this transitory part by random drawings from a distribution having mean zero and variance σ_b^2, that is, $\{b_t\}$ is a white noise process.

Sticky Innovation Model. The evolving nontransitory part z_t represents innovations that enter the system from time to time and *get stuck* there. These ''sticky'' innovations can arise from a multitude of causes, such as wear, corrosion, and human miscommunication. Thus, a car tire hits a sharp stone and *from that point onward* the tread is slightly damaged; a tiny crater caused by corrosion appears on the surface of a driving shaft and *remains*

there; certain details in the standard procedure for taking blood pressure in a hospital are forgotten and from that point on *permanently omitted or changed*. It is these nontransitory or sticky innovations that constitute the unwanted ''signal'' we wish to cancel out. Every system is subject to such influences. They continuously drive the increase in entropy if nothing is done to combat them. Such a sticky innovation model was suggested by Barnard (1959) and has a variogram that increases linearly with m. A special case of this model, which may also be used to approximate it, is the IMA(0, 1, 1) model:

$$N_t - N_{t-1} = a_t - \theta a_{t-1} \tag{15.2.10}$$

Recall, also, from Appendix A4.3 that if the nontransitory process Z_t is an IMA(0, 1, 1) process, then the disturbance process $N_t = b_t + z_t$ in (15.2.9) with the added white noise b_t, will again follow an IMA(0, 1, 1) model. Since for the IMA model (15.2.10) the EWMA of equation (15.2.1) with smoothing parameter θ provides a minimum mean square error (MMSE) forecast with forecast error $e_{t-1}(1) = a_t$, the corresponding discrete ''integral'' controller of (15.2.6) with $k_I = -\lambda/g$ produces MMSE control with $\varepsilon_t = a_t$. As we discuss later more formally, this is then a special case of the general MMSE linear feedback control scheme.

Dynamics. In discussion of the integral control scheme of equation (15.2.6), we assumed that any change made at the input of the system would have its full effect at the output in one time interval. The assumed dynamic equation for the response $\mathcal{Y}_t$ was, therefore,

$$\mathcal{Y}_t = gBX_{t+} \tag{15.2.11}$$

where we now denote the fixed level of the ''pulsed'' input X in the time interval from t until $t + 1$ by X_{t+}. A somewhat more general assumption is that the system can be described by the first-order difference equation

$$(1 + \xi\nabla)\mathcal{Y}_t = gBX_{t+} \tag{15.2.12}$$

(see, for example, (11.3.6)) or, equivalently,

$$(1 - \delta B)\mathcal{Y}_t = (1 - \delta)gBX_{t+} \qquad -1 < \delta < 1 \tag{15.2.13}$$

where $\xi = \delta/(1 - \delta)$ or, equivalently, $\delta = \xi/(1 + \xi)$. In that case at time $t + 1$ [cf. (15.1.1)], the deviation from target after adjustment is

$$\varepsilon_{t+1} = \mathcal{Y}_{t+1} + N_{t+1}$$

so that

$$\varepsilon_{t+1} = \frac{(1 - \delta)g}{1 - \delta B}X_{t+} + \hat{N}_t(1) + e_t(1)$$

where $\hat{N}_t(1)$ is some forecast of N_{t+1} made at time t with forecast error $e_t(1)$. Then, if we use the adjustment equation

$$X_{t+} - X_{t-1+} = x_t = -\frac{1 - \delta B}{(1 - \delta)g}[\hat{N}_t(1) - \hat{N}_{t-1}(1)]$$

the deviation ε_{t+1} from the target is equal to the forecast error $e_t(1)$. Thus, again we substitute the error in *forecasting* N_{t+1} for the deviation N_{t+1} itself. In particular, if $\hat{N}_t(1)$