Expression [4.1.1] is known as the *mean squared error* associated with the forecast $Y^*_{t+1|t}$, denoted

$$MSE(Y^*_{t+1|t}) \equiv E(Y_{t+1} - Y^*_{t+1|t})^2.$$

The forecast with the smallest mean squared error turns out to be the expectation of Y_{t+1} conditional on $\mathbf{X}_t$:

$$Y^*_{t+1|t} = E(Y_{t+1}|\mathbf{X}_t). \qquad [4.1.2]$$

To verify this claim, consider basing $Y^*_{t+1|t}$ on any function $g(\mathbf{X}_t)$ other than the conditional expectation,

$$Y^*_{t+1|t} = g(\mathbf{X}_t). \qquad [4.1.3]$$

For this candidate forecasting rule, the *MSE* would be

$$\begin{aligned} E[Y_{t+1} - g(\mathbf{X}_t)]^2 &= E[Y_{t+1} - E(Y_{t+1}|\mathbf{X}_t) + E(Y_{t+1}|\mathbf{X}_t) - g(\mathbf{X}_t)]^2 \\ &= E[Y_{t+1} - E(Y_{t+1}|\mathbf{X}_t)]^2 \\ &\quad + 2E\{[Y_{t+1} - E(Y_{t+1}|\mathbf{X}_t)][E(Y_{t+1}|\mathbf{X}_t) - g(\mathbf{X}_t)]\} \\ &\quad + E\{[E(Y_{t+1}|\mathbf{X}_t) - g(\mathbf{X}_t)]^2\}. \end{aligned} \qquad [4.1.4]$$

Write the middle term on the right side of [4.1.4] as

$$2E[\eta_{t+1}], \qquad [4.1.5]$$

where

$$\eta_{t+1} \equiv \{[Y_{t+1} - E(Y_{t+1}|\mathbf{X}_t)][E(Y_{t+1}|\mathbf{X}_t) - g(\mathbf{X}_t)]\}.$$

Consider first the expectation of η_{t+1} conditional on $\mathbf{X}_t$. Conditional on $\mathbf{X}_t$, the terms $E(Y_{t+1}|\mathbf{X}_t)$ and $g(\mathbf{X}_t)$ are known constants and can be factored out of this expectation:[1]

$$\begin{aligned} E[\eta_{t+1}|\mathbf{X}_t] &= [E(Y_{t+1}|\mathbf{X}_t) - g(\mathbf{X}_t)] \times E([Y_{t+1} - E(Y_{t+1}|\mathbf{X}_t)]|\mathbf{X}_t) \\ &= [E(Y_{t+1}|\mathbf{X}_t) - g(\mathbf{X}_t)] \times 0 \\ &= 0. \end{aligned}$$

By a straightforward application of the law of iterated expectations, equation [A.5.10], it follows that

$$E[\eta_{t+1}] = E_{\mathbf{X}_t}(E[\eta_{t+1}|\mathbf{X}_t]) = 0.$$

Substituting this back into [4.1.4] gives

$$E[Y_{t+1} - g(\mathbf{X}_t)]^2 = E[Y_{t+1} - E(Y_{t+1}|\mathbf{X}_t)]^2 + E([E(Y_{t+1}|\mathbf{X}_t) - g(\mathbf{X}_t)]^2). \qquad [4.1.6]$$

The second term on the right side of [4.1.6] cannot be made smaller than zero, and the first term does not depend on $g(\mathbf{X}_t)$. The function $g(\mathbf{X}_t)$ that makes the mean squared error [4.1.6] as small as possible is the function that sets the second term in [4.1.6] to zero:

$$E(Y_{t+1}|\mathbf{X}_t) = g(\mathbf{X}_t). \qquad [4.1.7]$$

Thus the forecast $g(\mathbf{X}_t)$ that minimizes the mean squared error is the conditional expectation $E(Y_{t+1}|\mathbf{X}_t)$, as claimed.

The *MSE* of this optimal forecast is

$$E[Y_{t+1} - g(\mathbf{X}_t)]^2 = E[Y_{t+1} - E(Y_{t+1}|\mathbf{X}_t)]^2. \qquad [4.1.8]$$

[1]The conditional expectation $E(Y_{t+1}|\mathbf{X}_t)$ represents the conditional population moment of the random variable Y_{t+1} and is not a function of the random variable Y_{t+1} itself. For example, if $Y_{t+1}|\mathbf{X}_t \sim N(\boldsymbol{\alpha}'\mathbf{X}_t, \boldsymbol{\Omega})$, then $E(Y_{t+1}|\mathbf{X}_t) = \boldsymbol{\alpha}'\mathbf{X}_t$, which does not depend on Y_{t+1}.

Forecasts Based on Linear Projection

We now restrict the class of forecasts considered by requiring the forecast $Y^*_{t+1|t}$ to be a linear function of $\mathbf{X}_t$:

$$Y^*_{t+1|t} = \boldsymbol{\alpha}'\mathbf{X}_t. \tag{4.1.9}$$

Suppose we were to find a value for $\boldsymbol{\alpha}$ such that the forecast error $(Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t)$ is uncorrelated with $\mathbf{X}_t$:

$$E[(Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t)\mathbf{X}_t'] = \mathbf{0}'. \tag{4.1.10}$$

If [4.1.10] holds, then the forecast $\boldsymbol{\alpha}'\mathbf{X}_t$ is called the *linear projection* of Y_{t+1} on $\mathbf{X}_t$.

The linear projection turns out to produce the smallest mean squared error among the class of linear forecasting rules. The proof of this claim closely parallels the demonstration of the optimality of the conditional expectation among the set of all possible forecasts. Let $\mathbf{g}'\mathbf{X}_t$ denote any arbitrary linear forecasting rule. Note that its *MSE* is

$$\begin{aligned} &E[Y_{t+1} - \mathbf{g}'\mathbf{X}_t]^2 \\ &\quad = E[Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t + \boldsymbol{\alpha}'\mathbf{X}_t - \mathbf{g}'\mathbf{X}_t]^2 \\ &\quad = E[Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t]^2 + 2E\{[Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t][\boldsymbol{\alpha}'\mathbf{X}_t - \mathbf{g}'\mathbf{X}_t]\} \\ &\qquad + E[\boldsymbol{\alpha}'\mathbf{X}_t - \mathbf{g}'\mathbf{X}_t]^2. \end{aligned} \tag{4.1.11}$$

As in the case of [4.1.4], the middle term on the right side of [4.1.11] is zero:

$$E([Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t][\boldsymbol{\alpha}'\mathbf{X}_t - \mathbf{g}'\mathbf{X}_t]) = (E[Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t]\mathbf{X}_t')[\boldsymbol{\alpha} - \mathbf{g}] = \mathbf{0}'[\boldsymbol{\alpha} - \mathbf{g}],$$

by virtue of [4.1.10]. Thus [4.1.11] simplifies to

$$E[Y_{t+1} - \mathbf{g}'\mathbf{X}_t]^2 = E[Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t]^2 + E[\boldsymbol{\alpha}'\mathbf{X}_t - \mathbf{g}'\mathbf{X}_t]^2. \tag{4.1.12}$$

The optimal linear forecast $\mathbf{g}'\mathbf{X}_t$ is the value that sets the second term in [4.1.12] equal to zero:

$$\mathbf{g}'\mathbf{X}_t = \boldsymbol{\alpha}'\mathbf{X}_t,$$

where $\boldsymbol{\alpha}'\mathbf{X}_t$ satisfies [4.1.10].

For $\boldsymbol{\alpha}'\mathbf{X}_t$ satisfying [4.1.10], we will use the notation

$$\hat{P}(Y_{t+1}|\mathbf{X}_t) = \boldsymbol{\alpha}'\mathbf{X}_t,$$

or sometimes simply

$$\hat{Y}_{t+1|t} = \boldsymbol{\alpha}'\mathbf{X}_t,$$

to indicate the linear projection of Y_{t+1} on $\mathbf{X}_t$. Notice that

$$MSE[\hat{P}(Y_{t+1}|\mathbf{X}_t)] \geq MSE[E(Y_{t+1}|\mathbf{X}_t)],$$

since the conditional expectation offers the best possible forecast.

For most applications a constant term will be included in the projection. We will use the symbol $\hat{E}$ to indicate a linear projection on a vector of random variables $\mathbf{X}_t$ along with a constant term:

$$\hat{E}(Y_{t+1}|\mathbf{X}_t) \equiv \hat{P}(Y_{t+1}|1, \mathbf{X}_t).$$

Properties of Linear Projection

It is straightforward to use [4.1.10] to calculate the projection coefficient $\boldsymbol{\alpha}$ in terms of the moments of Y_{t+1} and $\mathbf{X}_t$:

$$E(Y_{t+1}\mathbf{X}_t') = \boldsymbol{\alpha}'E(\mathbf{X}_t\mathbf{X}_t'),$$

or

$$\boldsymbol{\alpha}' = E(Y_{t+1}\mathbf{X}_t')[E(\mathbf{X}_t\mathbf{X}_t')]^{-1}, \qquad [4.1.13]$$

assuming that $E(\mathbf{X}_t\mathbf{X}_t')$ is a nonsingular matrix. When $E(\mathbf{X}_t\mathbf{X}_t')$ is singular, the coefficient vector $\boldsymbol{\alpha}$ is not uniquely determined by [4.1.10], though the product of this vector with the explanatory variables, $\boldsymbol{\alpha}'\mathbf{X}_t$, is uniquely determined by [4.1.10].[2]

The *MSE* associated with a linear projection is given by

$$E(Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t)^2 = E(Y_{t+1})^2 - 2E(\boldsymbol{\alpha}'\mathbf{X}_tY_{t+1}) + E(\boldsymbol{\alpha}'\mathbf{X}_t\mathbf{X}_t'\boldsymbol{\alpha}). \qquad [4.1.14]$$

Substituting [4.1.13] into [4.1.14] produces

$$\begin{aligned} E(Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t)^2 &= E(Y_{t+1})^2 - 2E(Y_{t+1}\mathbf{X}_t')[E(\mathbf{X}_t\mathbf{X}_t')]^{-1}E(\mathbf{X}_tY_{t+1}) \\ &\quad + E(Y_{t+1}\mathbf{X}_t')[E(\mathbf{X}_t\mathbf{X}_t')]^{-1} \\ &\quad \times E(\mathbf{X}_t\mathbf{X}_t')[E(\mathbf{X}_t\mathbf{X}_t')]^{-1}E(\mathbf{X}_tY_{t+1}) \\ &= E(Y_{t+1})^2 - E(Y_{t+1}\mathbf{X}_t')[E(\mathbf{X}_t\mathbf{X}_t')]^{-1}E(\mathbf{X}_tY_{t+1}). \end{aligned} \qquad [4.1.15]$$

Notice that if $\mathbf{X}_t$ includes a constant term, then the projection of $(aY_{t+1}+b)$ on $\mathbf{X}_t$ (where a and b are deterministic constants) is equal to

$$\hat{P}[(aY_{t+1} + b)|\mathbf{X}_t] = a\cdot\hat{P}(Y_{t+1}|\mathbf{X}_t) + b.$$

To see this, observe that $a\cdot\hat{P}(Y_{t+1}|\mathbf{X}_t) + b$ is a linear function of $\mathbf{X}_t$. Moreover, the forecast error,

$$[aY_{t+1} + b] - [a\cdot\hat{P}(Y_{t+1}|\mathbf{X}_t) + b] = a[Y_{t+1} - \hat{P}(Y_{t+1}|\mathbf{X}_t)],$$

is uncorrelated with $\mathbf{X}_t$, as required of a linear projection.

Linear Projection and Ordinary Least Squares Regression

Linear projection is closely related to ordinary least squares regression. This subsection discusses the relationship between the two concepts.

A linear regression model relates an observation on y_{t+1} to $\mathbf{x}_t$:

$$y_{t+1} = \boldsymbol{\beta}'\mathbf{x}_t + u_t. \qquad [4.1.16]$$

Given a sample of T observations on y and $\mathbf{x}$, the sample sum of squared residuals is defined as

$$\sum_{t=1}^{T} (y_{t+1} - \boldsymbol{\beta}'\mathbf{x}_t)^2. \qquad [4.1.17]$$

The value of $\boldsymbol{\beta}$ that minimizes [4.1.17], denoted $\mathbf{b}$, is the *ordinary least squares* (*OLS*) estimate of $\boldsymbol{\beta}$. The formula for $\mathbf{b}$ turns out to be

$$\mathbf{b} = \left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_ty_{t+1}\right], \qquad [4.1.18]$$

[2] If $E(\mathbf{X}_t\mathbf{X}_t')$ is singular, there exists a nonzero vector $\mathbf{c}$ such that $\mathbf{c}'\cdot E(\mathbf{X}_t\mathbf{X}_t')\cdot\mathbf{c} = E(\mathbf{c}'\mathbf{X}_t)^2 = 0$, so that some linear combination $\mathbf{c}'\mathbf{X}_t$ is equal to zero for all realizations. For example, if $\mathbf{X}_t$ consists of two random variables, the second variable must be a rescaled version of the first: $X_{2t} = c\cdot X_{1t}$. One could simply drop the redundant variables from such a system and calculate the linear projection of Y_{t+1} on $\mathbf{X}_t^*$, where $\mathbf{X}_t^*$ is a vector consisting of the nonredundant elements of $\mathbf{X}_t$. This linear projection $\boldsymbol{\alpha}^{*\prime}\mathbf{X}_t^*$ can be uniquely calculated from [4.1.13] with $\mathbf{X}_t$ in [4.1.13] replaced by $\mathbf{X}_t^*$. Any linear combination of the original variables $\boldsymbol{\alpha}'\mathbf{X}_t$ satisfying [4.1.10] represents this same random variable; that is, $\boldsymbol{\alpha}'\mathbf{X}_t = \boldsymbol{\alpha}^{*\prime}\mathbf{X}_t^*$ for all values of $\boldsymbol{\alpha}$ consistent with [4.1.10].

which equivalently can be written

$$\mathbf{b} = \left[(1/T) \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right]^{-1} \left[(1/T) \sum_{t=1}^{T} \mathbf{x}_t y_{t+1} \right]. \qquad [4.1.19]$$

Comparing the *OLS* coefficient estimate $\mathbf{b}$ in equation [4.1.19] with the linear projection coefficient $\boldsymbol{\alpha}$ in equation [4.1.13], we see that $\mathbf{b}$ is constructed from the sample moments $(1/T)\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$ and $(1/T)\Sigma_{t=1}^{T}\mathbf{x}_t y_{t+1}$ while $\boldsymbol{\alpha}$ is constructed from population moments $E(\mathbf{X}_t\mathbf{X}_t')$ and $E(\mathbf{X}_t Y_{t+1})$. Thus *OLS* regression is a summary of the particular sample observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ and $(y_2, y_3, \ldots, y_{T+1})$, whereas linear projection is a summary of the population characteristics of the stochastic process $\{\mathbf{X}_t, Y_{t+1}\}_{t=-\infty}^{\infty}$.

Although linear projection describes population moments and ordinary least squares describes sample moments, there is a formal mathematical sense in which the two operations are the same. Appendix 4.A to this chapter discusses this parallel and shows how the formulas for an *OLS* regression can be viewed as a special case of the formulas for a linear projection.

Notice that if the stochastic process $\{\mathbf{X}_t, Y_{t+1}\}$ is covariance-stationary and ergodic for second moments, then the sample moments will converge to the population moments as the sample size T goes to infinity:

$$(1/T) \sum_{t=1}^{T} \mathbf{X}_t\mathbf{X}_t' \xrightarrow{p} E(\mathbf{X}_t\mathbf{X}_t')$$

$$(1/T) \sum_{t=1}^{T} \mathbf{X}_t Y_{t+1} \xrightarrow{p} E(\mathbf{X}_t Y_{t+1}),$$

implying

$$\mathbf{b} \xrightarrow{p} \boldsymbol{\alpha}. \qquad [4.1.20]$$

Thus *OLS* regression of y_{t+1} on $\mathbf{x}_t$ yields a consistent estimate of the linear projection coefficient. Note that this result requires only that the process be ergodic for second moments. By contrast, structural econometric analysis requires much stronger assumptions about the relation between $\mathbf{X}$ and Y. The difference arises because structural analysis seeks the *effect* of $\mathbf{X}$ on Y. In structural analysis, changes in $\mathbf{X}$ are associated with a particular structural event such as a change in Federal Reserve policy, and the objective is to evaluate the consequences for Y. Where that is the objective, it is very important to consider the nature of the correlation between $\mathbf{X}$ and Y before relying on *OLS* estimates. In the case of linear projection, however, the only concern is forecasting, for which it does not matter whether it is $\mathbf{X}$ that causes Y or Y that causes $\mathbf{X}$. Their observed historical comovements (as summarized by $E(\mathbf{X}_t Y_{t+1})$) are all that is needed for calculating a forecast. Result [4.1.20] shows that ordinary least squares regression provides a sound basis for forecasting under very mild assumptions.

One possible violation of these assumptions should nevertheless be noted. Result [4.1.20] was derived by assuming a covariance-stationary, ergodic process. However, the moments of the data may have changed over time in fundamental ways, or the future environment may be different from that in the past. Where this is the case, ordinary least squares may be undesirable, and better forecasts can emerge from careful structural analysis.

Forecasting Vectors

The preceding results can be extended to forecast an $(n \times 1)$ vector $\mathbf{Y}_{t+1}$ on the basis of a linear function of an $(m \times 1)$ vector $\mathbf{X}_t$:

$$\hat{P}(\mathbf{Y}_{t+1}|\mathbf{X}_t) = \boldsymbol{\alpha}'\mathbf{X}_t \equiv \hat{\mathbf{Y}}_{t+1|t}. \qquad [4.1.21]$$

Then $\boldsymbol{\alpha}'$ would denote an $(n \times m)$ matrix of projection coefficients satisfying

$$E[(\mathbf{Y}_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t)\mathbf{X}_t'] = \mathbf{0}; \qquad [4.1.22]$$

that is, each of the n elements of $(\mathbf{Y}_{t+1} - \hat{\mathbf{Y}}_{t+1|t})$ is uncorrelated with each of the m elements of $\mathbf{X}_t$. Accordingly, the jth element of the vector $\hat{\mathbf{Y}}_{t+1|t}$ gives the minimum *MSE* forecast of the scalar $Y_{j,t+1}$. Moreover, to forecast any linear combination of the elements of $\mathbf{Y}_{t+1}$, say, $z_{t+1} = \mathbf{h}'\mathbf{Y}_{t+1}$, the minimum *MSE* forecast of z_{t+1} requires $(z_{t+1} - \hat{z}_{t+1|t})$ to be uncorrelated with $\mathbf{X}_t$. But since each of the elements of $(\mathbf{Y}_{t+1} - \hat{\mathbf{Y}}_{t+1|t})$ is uncorrelated with $\mathbf{X}_t$, clearly $\mathbf{h}'(\mathbf{Y}_{t+1} - \hat{\mathbf{Y}}_{t+1|t})$ is also uncorrelated with $\mathbf{X}_t$. Thus when $\hat{\mathbf{Y}}_{t+1|t}$ satisfies [4.1.22], then $\mathbf{h}'\hat{\mathbf{Y}}_{t+1|t}$ is the minimum *MSE* forecast of $\mathbf{h}'\mathbf{Y}_{t+1}$ for any value of $\mathbf{h}$.

From [4.1.22], the matrix of projection coefficients is given by

$$\boldsymbol{\alpha}' = [E(\mathbf{Y}_{t+1}\mathbf{X}_t')] \cdot [E(\mathbf{X}_t\mathbf{X}_t')]^{-1}. \qquad [4.1.23]$$

The matrix generalization of the formula for the mean squared error [4.1.15] is

$$\begin{aligned} MSE(\boldsymbol{\alpha}'\mathbf{X}_t) &\equiv E\{[\mathbf{Y}_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t] \cdot [\mathbf{Y}_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t]'\} \\ &= E(\mathbf{Y}_{t+1}\mathbf{Y}_{t+1}') - [E(\mathbf{Y}_{t+1}\mathbf{X}_t')] \cdot [E(\mathbf{X}_t\mathbf{X}_t')]^{-1} \cdot [E(\mathbf{X}_t\mathbf{Y}_{t+1}')]. \end{aligned} \qquad [4.1.24]$$

4.2. *Forecasts Based on an Infinite Number of Observations*

Forecasting Based on Lagged ε's

Consider a process with an $MA(\infty)$ representation

$$(Y_t - \mu) = \psi(L)\varepsilon_t \qquad [4.2.1]$$

with ε_t white noise and

$$\begin{aligned} \psi(L) &\equiv \sum_{j=0}^{\infty} \psi_j L^j \\ \psi_0 &= 1 \\ \sum_{j=0}^{\infty} |\psi_j| &< \infty. \end{aligned} \qquad [4.2.2]$$

Suppose that we have an infinite number of observations on ε through date t, $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$, and further know the values of μ and $\{\psi_1, \psi_2, \ldots\}$. Say we want to forecast the value of Y_{t+s}, that is, the value that Y will take on s periods from now. Note that [4.2.1] implies

$$\begin{aligned} Y_{t+s} = \mu + \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \cdots + \psi_{s-1}\varepsilon_{t+1} + \psi_s\varepsilon_t \\ + \psi_{s+1}\varepsilon_{t-1} + \cdots. \end{aligned} \qquad [4.2.3]$$

The optimal linear forecast takes the form

$$\hat{E}[Y_{t+s}|\varepsilon_t, \varepsilon_{t-1}, \ldots] = \mu + \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \psi_{s+2}\varepsilon_{t-2} + \cdots. \qquad [4.2.4]$$

That is, the unknown future ε's are set to their expected value of zero. The error associated with this forecast is

$$Y_{t+s} - \hat{E}[Y_{t+s}|\varepsilon_t, \varepsilon_{t-1}, \ldots] = \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \cdots + \psi_{s-1}\varepsilon_{t+1}. \quad [4.2.5]$$

In order for [4.2.4] to be the optimal linear forecast, condition [4.1.10] requires the forecast error to have mean zero and to be uncorrelated with ε_t, ε_{t-1}, ... It is readily confirmed that the error in [4.2.5] has these properties, so [4.2.4] must indeed be the linear projection, as claimed. The mean squared error associated with this forecast is

$$E(Y_{t+s} - \hat{E}[Y_{t+s}|\varepsilon_t, \varepsilon_{t-1}, \ldots])^2 = (1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{s-1}^2)\sigma^2. \quad [4.2.6]$$

For example, for an $MA(q)$ process,

$$\psi(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q,$$

the optimal linear forecast is

$$\hat{E}[Y_{t+s}|\varepsilon_t, \varepsilon_{t-1}, \ldots] \quad [4.2.7]$$
$$= \begin{cases} \mu + \theta_s\varepsilon_t + \theta_{s+1}\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q+s} & \text{for } s = 1, 2, \ldots, q \\ \mu & \text{for } s = q+1, q+2, \ldots. \end{cases}$$

The MSE is

$$\begin{cases} \sigma^2 & \text{for } s = 1 \\ (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_{s-1}^2)\sigma^2 & \text{for } s = 2, 3, \ldots, q \\ (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma^2 & \text{for } s = q+1, q+2, \ldots. \end{cases}$$

The MSE increases with the forecast horizon s up until $s = q$. If we try to forecast an $MA(q)$ farther than q periods into the future, the forecast is simply the unconditional mean of the series ($E(Y_t) = \mu$) and the MSE is the unconditional variance of the series ($\text{Var}(Y_t) = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma^2$).

These properties also characterize the $MA(\infty)$ case as the forecast horizon s goes to infinity. It is straightforward to establish from [4.2.2] that as $s \to \infty$, the forecast in [4.2.4] converges in mean square to μ, the unconditional mean. The MSE [4.2.6] likewise converges to $\sigma^2 \sum_{j=0}^{\infty}\psi_j^2$, which is the unconditional variance of the $MA(\infty)$ process [4.2.1].

A compact lag operator expression for the forecast in [4.2.4] is sometimes used. Consider taking the polynomial $\psi(L)$ and dividing by L^s:

$$\frac{\psi(L)}{L^s} = L^{-s} + \psi_1 L^{1-s} + \psi_2 L^{2-s} + \cdots + \psi_{s-1}L^{-1} + \psi_s L^0 + \psi_{s+1}L^1 + \psi_{s+2}L^2 + \cdots.$$

The *annihilation operator*[3] (indicated by $[\cdot]_+$) replaces negative powers of L by zero; for example,

$$\left[\frac{\psi(L)}{L^s}\right]_+ = \psi_s + \psi_{s+1}L^1 + \psi_{s+2}L^2 + \cdots. \quad [4.2.8]$$

Comparing [4.2.8] with [4.2.4], the optimal forecast could be written in lag operator notation as

$$\hat{E}[Y_{t+s}|\varepsilon_t, \varepsilon_{t-1}, \ldots] = \mu + \left[\frac{\psi(L)}{L^s}\right]_+ \varepsilon_t. \quad [4.2.9]$$

[3]This discussion of forecasting based on the annihilation operator is similar to that in Sargent (1987).

Forecasting Based on Lagged Y's

The previous forecasts were based on the assumption that ε_t is observed directly. In the usual forecasting situation, we actually have observations on lagged Y's, not lagged ε's. Suppose that the process [4.2.1] has an $AR(\infty)$ representation given by

$$\eta(L)(Y_t - \mu) = \varepsilon_t, \qquad [4.2.10]$$

where $\eta(L) \equiv \sum_{j=0}^{\infty}\eta_j L^j$, $\eta_0 = 1$, and $\sum_{j=0}^{\infty}|\eta_j| < \infty$. Suppose further that the AR polynomial $\eta(L)$ and the MA polynomial $\psi(L)$ are related by

$$\eta(L) = [\psi(L)]^{-1}. \qquad [4.2.11]$$

A covariance-stationary $AR(p)$ model of the form

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)(Y_t - \mu) = \varepsilon_t, \qquad [4.2.12]$$

or, more compactly,

$$\phi(L)(Y_t - \mu) = \varepsilon_t,$$

clearly satisfies these requirements, with $\eta(L) = \phi(L)$ and $\psi(L) = [\phi(L)]^{-1}$. An $MA(q)$ process

$$Y_t - \mu = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t \qquad [4.2.13]$$

or

$$Y_t - \mu = \theta(L)\varepsilon_t$$

is also of this form, with $\psi(L) = \theta(L)$ and $\eta(L) = [\theta(L)]^{-1}$, provided that [4.2.13] is based on the invertible representation. With a noninvertible $MA(q)$, the roots must first be flipped as described in Section 3.7 before applying the formulas given in this section. An $ARMA(p, q)$ also satisfies [4.2.10] and [4.2.11] with $\psi(L) = \theta(L)/\phi(L)$, provided that the autoregressive operator $\phi(L)$ satisfies the stationarity condition (roots of $\phi(z) = 0$ lie outside the unit circle) and that the moving average operator $\theta(L)$ satisfies the invertibility condition (roots of $\theta(z) = 0$ lie outside the unit circle).

Where the restrictions associated with [4.2.10] and [4.2.11] are satisfied, observations on $\{Y_t, Y_{t-1}, \ldots\}$ will be sufficient to construct $\{\varepsilon_t, \varepsilon_{t-1}, \ldots\}$. For example, for an $AR(1)$ process [4.2.10] would be

$$(1 - \phi L)(Y_t - \mu) = \varepsilon_t. \qquad [4.2.14]$$

Thus, given ϕ and μ and observation of Y_t and Y_{t-1}, the value of ε_t can be constructed from

$$\varepsilon_t = (Y_t - \mu) - \phi(Y_{t-1} - \mu).$$

For an $MA(1)$ process written in invertible form, [4.2.10] would be

$$(1 + \theta L)^{-1}(Y_t - \mu) = \varepsilon_t.$$

Given an infinite number of observations on Y, we could construct ε from

$$\begin{aligned}\varepsilon_t = (Y_t - \mu) - \theta(Y_{t-1} - \mu) + \theta^2(Y_{t-2} - \mu) \\ - \theta^3(Y_{t-3} - \mu) + \cdots.\end{aligned} \qquad [4.2.15]$$

Under these conditions, [4.2.10] can be substituted into [4.2.9] to obtain the forecast of Y_{t+s} as a function of lagged Y's:

$$\hat{E}[Y_{t+s}|Y_t, Y_{t-1}, \ldots] = \mu + \left[\frac{\psi(L)}{L^s}\right]_+ \eta(L)(Y_t - \mu);$$

or, using [4.2.11],

$$\hat{E}[Y_{t+s}|Y_t, Y_{t-1}, \ldots] = \mu + \left[\frac{\psi(L)}{L^s}\right]_+ \frac{1}{\psi(L)}(Y_t - \mu). \qquad [4.2.16]$$

Equation [4.2.16] is known as the *Wiener-Kolmogorov prediction formula*. Several examples of using this forecasting rule follow.

Forecasting an AR(1) *Process*

For the covariance-stationary $AR(1)$ process [4.2.14], we have

$$\psi(L) = 1/(1 - \phi L) = 1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots \qquad [4.2.17]$$

and

$$\left[\frac{\psi(L)}{L^s}\right]_+ = \phi^s + \phi^{s+1}L^1 + \phi^{s+2}L^2 + \cdots = \phi^s/(1 - \phi L). \qquad [4.2.18]$$

Substituting [4.2.18] into [4.2.16] yields the optimal linear s-period-ahead forecast for a stationary $AR(1)$ process:

$$\begin{aligned}\hat{E}[Y_{t+s}|Y_t, Y_{t-1}, \ldots] &= \mu + \frac{\phi^s}{1 - \phi L}(1 - \phi L)(Y_t - \mu) \\ &= \mu + \phi^s(Y_t - \mu).\end{aligned} \qquad [4.2.19]$$

The forecast decays geometrically from $(Y_t - \mu)$ toward μ as the forecast horizon s increases. From [4.2.17], the moving average weight ψ_j is given by ϕ^j, so from [4.2.6], the mean squared s-period-ahead forecast error is

$$[1 + \phi^2 + \phi^4 + \cdots + \phi^{2(s-1)}]\sigma^2.$$

Notice that this grows with s and asymptotically approaches $\sigma^2/(1 - \phi^2)$, the unconditional variance of Y.

Forecasting an AR(p) *Process*

Next consider forecasting the stationary $AR(p)$ process [4.2.12]. The Wiener-Kolmogorov formula in [4.2.16] essentially expresses the value of $(Y_{t+s} - \mu)$ in terms of initial values $\{(Y_t - \mu), (Y_{t-1} - \mu), \ldots\}$ and subsequent values of $\{\varepsilon_{t+1}, \varepsilon_{t+2}, \ldots, \varepsilon_{t+s}\}$ and then drops the terms involving future ε's. An expression of this form was provided by equation [1.2.26], which described the value of a variable subject to a pth-order difference equation in terms of initial conditions and subsequent shocks:

$$\begin{aligned}Y_{t+s} - \mu = f_{11}^{(s)}(Y_t - \mu) + f_{12}^{(s)}(Y_{t-1} - \mu) + \cdots + f_{1p}^{(s)}(Y_{t-p+1} - \mu) \\ + \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \psi_2\varepsilon_{t+s-2} + \cdots + \psi_{s-1}\varepsilon_{t+1},\end{aligned} \qquad [4.2.20]$$

where

$$\psi_j = f_{11}^{(j)}. \qquad [4.2.21]$$

Recall that $f_{11}^{(j)}$ denotes the (1, 1) element of $\mathbf{F}^j$, $f_{12}^{(j)}$ denotes the (1, 2) element of $\mathbf{F}^j$, and so on, where $\mathbf{F}$ is the following $(p \times p)$ matrix:

$$\mathbf{F} \equiv \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

The optimal s-period-ahead forecast is thus

$$\begin{aligned} \hat{Y}_{t+s|t} = \mu &+ f_{11}^{(s)}(Y_t - \mu) + f_{12}^{(s)}(Y_{t-1} - \mu) + \cdots \\ &+ f_{1p}^{(s)}(Y_{t-p+1} - \mu). \end{aligned} \tag{4.2.22}$$

Notice that for any forecast horizon s the optimal forecast is a constant plus a linear function of $\{Y_t, Y_{t-1}, \ldots, Y_{t-p+1}\}$. The associated forecast error is

$$Y_{t+s} - \hat{Y}_{t+s|t} = \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \psi_2\varepsilon_{t+s-2} + \cdots + \psi_{s-1}\varepsilon_{t+1}. \tag{4.2.23}$$

The easiest way to calculate the forecast in [4.2.22] is through a simple recursion. This recursion can be deduced independently from a principle known as the *law of iterated projections*, which will be proved formally in Section 4.5. Suppose that at date t we wanted to make a one-period-ahead forecast of Y_{t+1}. The optimal forecast is clearly

$$\begin{aligned} (\hat{Y}_{t+1|t} - \mu) = \phi_1(Y_t - \mu) &+ \phi_2(Y_{t-1} - \mu) + \cdots \\ &+ \phi_p(Y_{t-p+1} - \mu). \end{aligned} \tag{4.2.24}$$

Consider next a two-period-ahead forecast. Suppose that at date $t + 1$ we were to make a one-period-ahead forecast of Y_{t+2}. Replacing t with $t + 1$ in [4.2.24] gives the optimal forecast as

$$\begin{aligned} (\hat{Y}_{t+2|t+1} - \mu) = \phi_1(Y_{t+1} - \mu) &+ \phi_2(Y_t - \mu) + \cdots \\ &+ \phi_p(Y_{t-p+2} - \mu). \end{aligned} \tag{4.2.25}$$

The law of iterated projections asserts that if this date $t + 1$ forecast of Y_{t+2} is projected on date t information, the result is the date t forecast of Y_{t+2}. At date t the values $Y_t, Y_{t-1}, \ldots, Y_{t-p+2}$ in [4.2.25] are known. Thus,

$$\begin{aligned} (\hat{Y}_{t+2|t} - \mu) = \phi_1(\hat{Y}_{t+1|t} - \mu) &+ \phi_2(Y_t - \mu) + \cdots \\ &+ \phi_p(Y_{t-p+2} - \mu). \end{aligned} \tag{4.2.26}$$

Substituting [4.2.24] into [4.2.26] then yields the two-period-ahead forecast for an $AR(p)$ process:

$$\begin{aligned} (\hat{Y}_{t+2|t} - \mu) &= \phi_1[\phi_1(Y_t - \mu) + \phi_2(Y_{t-1} - \mu) + \cdots + \phi_p(Y_{t-p+1} - \mu)] \\ &\quad + \phi_2(Y_t - \mu) + \phi_3(Y_{t-1} - \mu) + \cdots + \phi_p(Y_{t-p+2} - \mu) \\ &= (\phi_1^2 + \phi_2)(Y_t - \mu) + (\phi_1\phi_2 + \phi_3)(Y_{t-1} - \mu) + \cdots \\ &\quad + (\phi_1\phi_{p-1} + \phi_p)(Y_{t-p+2} - \mu) + \phi_1\phi_p(Y_{t-p+1} - \mu). \end{aligned}$$

The s-period-ahead forecasts of an $AR(p)$ process can be obtained by iterating on

$$\begin{aligned} (\hat{Y}_{t+j|t} - \mu) = \phi_1(\hat{Y}_{t+j-1|t} - \mu) &+ \phi_2(\hat{Y}_{t+j-2|t} - \mu) + \cdots \\ &+ \phi_p(\hat{Y}_{t+j-p|t} - \mu) \end{aligned} \tag{4.2.27}$$

for $j = 1, 2, \ldots, s$ where

$$\hat{Y}_{\tau|t} = Y_\tau \qquad \text{for } \tau \leq t.$$

Forecasting an MA*(1) Process*

Next consider an invertible $MA(1)$ representation,

$$Y_t - \mu = (1 + \theta L)\varepsilon_t \qquad [4.2.28]$$

with $|\theta| < 1$. Replacing $\psi(L)$ in the Wiener-Kolmogorov formula [4.2.16] with $(1 + \theta L)$ gives

$$\hat{Y}_{t+s|t} = \mu + \left[\frac{1 + \theta L}{L^s}\right]_+ \frac{1}{1 + \theta L}(Y_t - \mu). \qquad [4.2.29]$$

To forecast an $MA(1)$ process one period into the future ($s = 1$),

$$\left[\frac{1 + \theta L}{L^1}\right]_+ = \theta,$$

and so

$$\begin{aligned}\hat{Y}_{t+1|t} &= \mu + \frac{\theta}{1 + \theta L}(Y_t - \mu) \\ &= \mu + \theta(Y_t - \mu) - \theta^2(Y_{t-1} - \mu) + \theta^3(Y_{t-2} - \mu) - \cdots.\end{aligned} \qquad [4.2.30]$$

It is sometimes useful to write [4.2.28] as

$$\varepsilon_t = \frac{1}{1 + \theta L}(Y_t - \mu)$$

and view ε_t as the outcome of an infinite recursion,

$$\hat{\varepsilon}_t = (Y_t - \mu) - \theta\hat{\varepsilon}_{t-1}. \qquad [4.2.31]$$

The one-period-ahead forecast [4.2.30] could then be written as

$$\hat{Y}_{t+1|t} = \mu + \theta\hat{\varepsilon}_t. \qquad [4.2.32]$$

Equation [4.2.31] is in fact an exact characterization of ε_t, deduced from simple rearrangement of [4.2.28]. The "hat" notation ($\hat{\varepsilon}_t$) is introduced at this point in anticipation of the approximations to ε_t that will be introduced in the following section and substituted into [4.2.31] and [4.2.32].

To forecast an $MA(1)$ process for $s = 2, 3, \ldots$ periods into the future,

$$\left[\frac{1 + \theta L}{L^s}\right]_+ = 0 \qquad \text{for } s = 2, 3, \ldots;$$

and so, from [4.2.29],

$$\hat{Y}_{t+s|t} = \mu \qquad \text{for } s = 2, 3, \ldots. \qquad [4.2.33]$$

Forecasting an MA(q) *Process*

For an invertible $MA(q)$ process,

$$(Y_t - \mu) = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t,$$

the forecast [4.2.16] becomes

$$\hat{Y}_{t+s|t} = \mu + \left[\frac{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q}{L^s}\right]_+ \times \frac{1}{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q}(Y_t - \mu). \quad [4.2.34]$$

Now

$$\left[\frac{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q}{L^s}\right]_+ = \begin{cases} \theta_s + \theta_{s+1} L + \theta_{s+2} L^2 + \cdots + \theta_q L^{q-s} & \text{for } s = 1, 2, \ldots, q \\ 0 & \text{for } s = q+1, q+2, \ldots. \end{cases}$$

Thus, for horizons of $s = 1, 2, \ldots, q$, the forecast is given by

$$\hat{Y}_{t+s|t} = \mu + (\theta_s + \theta_{s+1} L + \theta_{s+2} L^2 + \cdots + \theta_q L^{q-s})\hat{\varepsilon}_t, \quad [4.2.35]$$

where $\hat{\varepsilon}_t$ can be characterized by the recursion

$$\hat{\varepsilon}_t = (Y_t - \mu) - \theta_1 \hat{\varepsilon}_{t-1} - \theta_2 \hat{\varepsilon}_{t-2} - \cdots - \theta_q \hat{\varepsilon}_{t-q}. \quad [4.2.36]$$

A forecast farther than q periods into the future is simply the unconditional mean μ.

Forecasting an ARMA*(1, 1) Process*

For an $ARMA(1, 1)$ process

$$(1 - \phi L)(Y_t - \mu) = (1 + \theta L)\varepsilon_t$$

that is stationary ($|\phi| < 1$) and invertible ($|\theta| < 1$),

$$\hat{Y}_{t+s|t} = \mu + \left[\frac{1 + \theta L}{(1 - \phi L)L^s}\right]_+ \frac{1 - \phi L}{1 + \theta L}(Y_t - \mu). \quad [4.2.37]$$

Here

$$\begin{aligned} &\left[\frac{1 + \theta L}{(1 - \phi L)L^s}\right]_+ \\ &= \left[\frac{(1 + \phi L + \phi^2 L^2 + \cdots)}{L^s} + \frac{\theta L(1 + \phi L + \phi^2 L^2 + \cdots)}{L^s}\right]_+ \\ &= (\phi^s + \phi^{s+1} L + \phi^{s+2} L^2 + \cdots) + \theta(\phi^{s-1} + \phi^s L + \phi^{s+1} L^2 + \cdots) \\ &= (\phi^s + \theta\phi^{s-1})(1 + \phi L + \phi^2 L^2 + \cdots) \\ &= \frac{\phi^s + \theta\phi^{s-1}}{1 - \phi L}. \end{aligned} \quad [4.2.38]$$

Substituting [4.2.38] into [4.2.37] gives

$$\begin{aligned} \hat{Y}_{t+s|t} &= \mu + \left[\frac{\phi^s + \theta\phi^{s-1}}{1 - \phi L}\right]\frac{1 - \phi L}{1 + \theta L}(Y_t - \mu) \\ &= \mu + \frac{\phi^s + \theta\phi^{s-1}}{1 + \theta L}(Y_t - \mu). \end{aligned} \quad [4.2.39]$$

Note that for $s = 2, 3, \ldots,$ the forecast [4.2.39] obeys the recursion

$$(\hat{Y}_{t+s|t} - \mu) = \phi(\hat{Y}_{t+s-1|t} - \mu).$$

Thus, beyond one period, the forecast decays geometrically at the rate ϕ toward the unconditional mean μ. The one-period-ahead forecast ($s = 1$) is given by

$$\hat{Y}_{t+1|t} = \mu + \frac{\phi + \theta}{1 + \theta L}(Y_t - \mu). \qquad [4.2.40]$$

This can equivalently be written

$$(\hat{Y}_{t+1|t} - \mu) = \frac{\phi(1 + \theta L) + \theta(1 - \phi L)}{1 + \theta L}(Y_t - \mu) = \phi(Y_t - \mu) + \theta\hat{\varepsilon}_t \qquad [4.2.41]$$

where

$$\hat{\varepsilon}_t = \frac{(1 - \phi L)}{(1 + \theta L)}(Y_t - \mu)$$

or

$$\hat{\varepsilon}_t = (Y_t - \mu) - \phi(Y_{t-1} - \mu) - \theta\hat{\varepsilon}_{t-1} = Y_t - \hat{Y}_{t|t-1}. \qquad [4.2.42]$$

Forecasting an ARMA(p, q) *Process*

Finally, consider forecasting a stationary and invertible $ARMA(p, q)$ process:

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)(Y_t - \mu) = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t.$$

The natural generalizations of [4.2.41] and [4.2.42] are

$$\begin{aligned}(\hat{Y}_{t+1|t} - \mu) = {} & \phi_1(Y_t - \mu) + \phi_2(Y_{t-1} - \mu) + \cdots \\ & + \phi_p(Y_{t-p+1} - \mu) + \theta_1\hat{\varepsilon}_t + \theta_2\hat{\varepsilon}_{t-1} + \cdots + \theta_q\hat{\varepsilon}_{t-q+1},\end{aligned} \qquad [4.2.43]$$

with $\{\hat{\varepsilon}_t\}$ generated recursively from

$$\hat{\varepsilon}_t = Y_t - \hat{Y}_{t|t-1}. \qquad [4.2.44]$$

The s-period-ahead forecasts would be

$$(\hat{Y}_{t+s|t} - \mu) \qquad [4.2.45]$$

$$= \begin{cases} \phi_1(\hat{Y}_{t+s-1|t} - \mu) + \phi_2(\hat{Y}_{t+s-2|t} - \mu) + \cdots + \phi_p(\hat{Y}_{t+s-p|t} - \mu) \\ \quad + \theta_s\hat{\varepsilon}_t + \theta_{s+1}\hat{\varepsilon}_{t-1} + \cdots + \theta_q\hat{\varepsilon}_{t+s-q} \qquad \text{for } s = 1, 2, \ldots, q \\ \phi_1(\hat{Y}_{t+s-1|t} - \mu) + \phi_2(\hat{Y}_{t+s-2|t} - \mu) + \cdots + \phi_p(\hat{Y}_{t+s-p|t} - \mu) \\ \qquad\qquad \text{for } s = q + 1, q + 2, \ldots, \end{cases}$$

where

$$\hat{Y}_{\tau|t} = Y_\tau \qquad \text{for } \tau \leq t.$$

Thus for a forecast horizon s greater than the moving average order q, the forecasts follow a pth-order difference equation governed solely by the autoregressive parameters.

4.3. *Forecasts Based on a Finite Number of Observations*

The formulas in the preceding section assumed that we had an infinite number of past observations on Y, $\{Y_t, Y_{t-1}, \ldots\}$, and knew with certainty population parameters such as μ, ϕ, and θ. This section continues to assume that population parameters are known with certainty, but develops forecasts based on a finite number of observations $\{Y_t, Y_{t-1}, \ldots, Y_{t-m+1}\}$.

For forecasting an $AR(p)$ process, an optimal s-period-ahead linear forecast based on an infinite number of observations $\{Y_t, Y_{t-1}, \ldots\}$ in fact makes use of only the p most recent values $\{Y_t, Y_{t-1}, \ldots, Y_{t-p+1}\}$. For an MA or $ARMA$ process, however, we would in principle require all of the historical values of Y in order to implement the formulas of the preceding section.

Approximations to Optimal Forecasts

One approach to forecasting based on a finite number of observations is to act as if presample ε's were all equal to zero. The idea is thus to use the approximation

$$\hat{E}(Y_{t+s}|Y_t, Y_{t-1}, \ldots) \cong \hat{E}(Y_{t+s}|Y_t, Y_{t-1}, \ldots, Y_{t-m+1}, \varepsilon_{t-m} = 0, \varepsilon_{t-m-1} = 0, \ldots). \quad [4.3.1]$$

For example, consider forecasting an $MA(q)$ process. The recursion [4.2.36] can be started by setting

$$\hat{\varepsilon}_{t-m} = \hat{\varepsilon}_{t-m-1} = \cdots = \hat{\varepsilon}_{t-m-q+1} = 0 \quad [4.3.2]$$

and then iterating on [4.2.36] to generate $\hat{\varepsilon}_{t-m+1}, \hat{\varepsilon}_{t-m+2}, \ldots, \hat{\varepsilon}_t$. These calculations produce

$$\begin{aligned}
\hat{\varepsilon}_{t-m+1} &= (Y_{t-m+1} - \mu),\\
\hat{\varepsilon}_{t-m+2} &= (Y_{t-m+2} - \mu) - \theta_1\hat{\varepsilon}_{t-m+1},\\
\hat{\varepsilon}_{t-m+3} &= (Y_{t-m+3} - \mu) - \theta_1\hat{\varepsilon}_{t-m+2} - \theta_2\hat{\varepsilon}_{t-m+1},
\end{aligned}$$

and so on. The resulting values for $(\hat{\varepsilon}_t, \hat{\varepsilon}_{t-1}, \ldots, \hat{\varepsilon}_{t-q+s})$ are then substituted directly into [4.2.35] to produce the forecast [4.3.1]. For example, for $s = q = 1$, the forecast would be

$$\begin{aligned}
\hat{Y}_{t+1|t} = \mu &+ \theta(Y_t - \mu) - \theta^2(Y_{t-1} - \mu)\\
&+ \theta^3(Y_{t-2} - \mu) - \cdots + (-1)^{m-1}\theta^m(Y_{t-m+1} - \mu),
\end{aligned} \quad [4.3.3]$$

which is to be used as an approximation to the $AR(\infty)$ forecast,

$$\mu + \theta(Y_t - \mu) - \theta^2(Y_{t-1} - \mu) + \theta^3(Y_{t-2} - \mu) - \cdots. \quad [4.3.4]$$

For m large and $|\theta|$ small, this clearly gives an excellent approximation. For $|\theta|$ closer to unity, the approximation may be poorer. Note that if the moving average operator is noninvertible, the forecast [4.3.1] is inappropriate and should not be used.

Exact Finite-Sample Forecasts

An alternative approach is to calculate the exact projection of Y_{t+1} on its m most recent values. Let

$$\mathbf{X}_t \equiv \begin{bmatrix} 1 \\ Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-m+1} \end{bmatrix}.$$

We thus seek a linear forecast of the form

$$\boldsymbol{\alpha}^{(m)\prime}\mathbf{X}_t = \alpha_0^{(m)} + \alpha_1^{(m)}Y_t + \alpha_2^{(m)}Y_{t-1} + \cdots + \alpha_m^{(m)}Y_{t-m+1}. \qquad [4.3.5]$$

The coefficient relating Y_{t+1} to Y_t in a projection of Y_{t+1} on the m most recent values of Y is denoted $\alpha_1^{(m)}$ in [4.3.5]. This will in general be different from the coefficient relating Y_{t+1} to Y_t in a projection of Y_{t+1} on the $m + 1$ most recent values of Y; the latter coefficient would be denoted $\alpha_1^{(m+1)}$.

If Y_t is covariance-stationary, then $E(Y_tY_{t-j}) = \gamma_j + \mu^2$. Setting $\mathbf{X}_t = (1, Y_t, Y_{t-1}, \ldots, Y_{t-m+1})'$ in [4.1.13] implies

$$\begin{aligned}\boldsymbol{\alpha}^{(m)\prime} &\equiv [\alpha_0^{(m)} \quad \alpha_1^{(m)} \quad \alpha_2^{(m)} \cdots \alpha_m^{(m)}] \\ &= [\mu \quad (\gamma_1 + \mu^2) \quad (\gamma_2 + \mu^2) \cdots (\gamma_m + \mu^2)] \\ &\quad \times \begin{bmatrix} 1 & \mu & \mu & \cdots & \mu \\ \mu & \gamma_0 + \mu^2 & \gamma_1 + \mu^2 & \cdots & \gamma_{m-1} + \mu^2 \\ \mu & \gamma_1 + \mu^2 & \gamma_0 + \mu^2 & \cdots & \gamma_{m-2} + \mu^2 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mu & \gamma_{m-1} + \mu^2 & \gamma_{m-2} + \mu^2 & \cdots & \gamma_0 + \mu^2 \end{bmatrix}^{-1}. \qquad [4.3.6]\end{aligned}$$

When a constant term is included in $\mathbf{X}_t$, it is more convenient to express variables in deviations from the mean. Then we could calculate the projection of $(Y_{t+1} - \mu)$ on $\mathbf{X}_t = [(Y_t - \mu), (Y_{t-1} - \mu), \ldots, (Y_{t-m+1} - \mu)]'$:

$$\begin{aligned}\hat{Y}_{t+1|t} - \mu = \alpha_1^{(m)}(Y_t - \mu) + \alpha_2^{(m)}(Y_{t-1} - \mu) + \cdots \\ + \alpha_m^{(m)}(Y_{t-m+1} - \mu).\end{aligned} \qquad [4.3.7]$$

For this definition of $\mathbf{X}_t$ the coefficients can be calculated directly from [4.1.13] to be

$$\begin{bmatrix} \alpha_1^{(m)} \\ \alpha_2^{(m)} \\ \vdots \\ \alpha_m^{(m)} \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{m-2} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{bmatrix}. \qquad [4.3.8]$$

We will demonstrate in Section 4.5 that the coefficients $(\alpha_1^{(m)}, \alpha_2^{(m)}, \ldots, \alpha_m^{(m)})$ in equations [4.3.8] and [4.3.6] are identical. This is analogous to a familiar result for ordinary least squares regression—slope coefficients would be unchanged if all variables are expressed in deviations from their sample means and the constant term is dropped from the regression.

To generate an s-period-ahead forecast $\hat{Y}_{t+s|t}$, we would use

$$\hat{Y}_{t+s|t} = \mu + \alpha_1^{(m,s)}(Y_t - \mu) + \alpha_2^{(m,s)}(Y_{t-1} - \mu) + \cdots + \alpha_m^{(m,s)}(Y_{t-m+1} - \mu),$$

where

$$\begin{bmatrix} \alpha_1^{(m,s)} \\ \alpha_2^{(m,s)} \\ \vdots \\ \alpha_m^{(m,s)} \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{m-2} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_s \\ \gamma_{s+1} \\ \vdots \\ \gamma_{s+m-1} \end{bmatrix}. \qquad [4.3.9]$$

Using expressions such as [4.3.8] requires inverting an $(m \times m)$ matrix. Several algorithms can be used to evaluate [4.3.8] using relatively simple calculations. One approach is based on the Kalman filter discussed in Chapter 13, which can generate exact finite-sample forecasts for a broad class of processes including any *ARMA* specification. A second approach is based on triangular factorization of the matrix in [4.3.8]. This second approach is developed in the next two sections. This approach will prove helpful for the immediate question of calculating finite-sample forecasts and is also a useful device for establishing a number of later results.

4.4. *The Triangular Factorization of a Positive Definite Symmetric Matrix*

Any positive definite symmetric $(n \times n)$ matrix $\mathbf{\Omega}$ has a unique representation of the form

$$\mathbf{\Omega} = \mathbf{ADA}', \qquad [4.4.1]$$

where $\mathbf{A}$ is a lower triangular matrix with 1s along the principal diagonal,

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{21} & 1 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & 1 \end{bmatrix},$$

and $\mathbf{D}$ is a diagonal matrix,

$$\mathbf{D} = \begin{bmatrix} d_{11} & 0 & 0 & \cdots & 0 \\ 0 & d_{22} & 0 & \cdots & 0 \\ 0 & 0 & d_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & d_{nn} \end{bmatrix},$$

where $d_{ii} > 0$ for all i. This is known as the *triangular factorization* of $\mathbf{\Omega}$.

To see how the triangular factorization can be calculated, consider

$$\boldsymbol{\Omega} = \begin{bmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} & \cdots & \Omega_{1n} \\ \Omega_{21} & \Omega_{22} & \Omega_{23} & \cdots & \Omega_{2n} \\ \Omega_{31} & \Omega_{32} & \Omega_{33} & \cdots & \Omega_{3n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \Omega_{n1} & \Omega_{n2} & \Omega_{n3} & \cdots & \Omega_{nn} \end{bmatrix}. \qquad [4.4.2]$$

We assume that $\boldsymbol{\Omega}$ is positive definite, meaning that $\mathbf{x}'\boldsymbol{\Omega}\mathbf{x} > 0$ for any nonzero $(n \times 1)$ vector $\mathbf{x}$. We also assume that $\boldsymbol{\Omega}$ is symmetric, so that $\Omega_{ij} = \Omega_{ji}$.

The matrix $\boldsymbol{\Omega}$ can be transformed into a matrix with zero in the (2, 1) position by multiplying the first row of $\boldsymbol{\Omega}$ by $\Omega_{21}\Omega_{11}^{-1}$ and subtracting the resulting row from the second. A zero can be put in the (3, 1) position by multiplying the first row by $\Omega_{31}\Omega_{11}^{-1}$ and subtracting the resulting row from the third. We proceed in this fashion down the first column. This set of operations can be summarized as premultiplying $\boldsymbol{\Omega}$ by the following matrix:

$$\mathbf{E}_1 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\Omega_{21}\Omega_{11}^{-1} & 1 & 0 & \cdots & 0 \\ -\Omega_{31}\Omega_{11}^{-1} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ -\Omega_{n1}\Omega_{11}^{-1} & 0 & 0 & \cdots & 1 \end{bmatrix}. \qquad [4.4.3]$$

This matrix always exists, provided that $\Omega_{11} \neq 0$. This is ensured in the present case, because Ω_{11} is equal to $\mathbf{e}_1'\boldsymbol{\Omega}\mathbf{e}_1$, where $\mathbf{e}_1' = [1 \quad 0 \quad 0 \cdots 0]$. Since $\boldsymbol{\Omega}$ is positive definite, $\mathbf{e}_1'\boldsymbol{\Omega}\mathbf{e}_1$ must be greater than zero.

When $\boldsymbol{\Omega}$ is premultiplied by $\mathbf{E}_1$ and postmultiplied by $\mathbf{E}_1'$ the result is

$$\mathbf{E}_1\boldsymbol{\Omega}\mathbf{E}_1' = \mathbf{H}, \qquad [4.4.4]$$

where

$$\mathbf{H} = \begin{bmatrix} h_{11} & 0 & 0 & \cdots & 0 \\ 0 & h_{22} & h_{23} & \cdots & h_{2n} \\ 0 & h_{32} & h_{33} & \cdots & h_{3n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & h_{n2} & h_{n3} & \cdots & h_{nn} \end{bmatrix} \qquad [4.4.5]$$

$$= \begin{bmatrix} \Omega_{11} & 0 & 0 & \cdots & 0 \\ 0 & \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} & \Omega_{23} - \Omega_{21}\Omega_{11}^{-1}\Omega_{13} & \cdots & \Omega_{2n} - \Omega_{21}\Omega_{11}^{-1}\Omega_{1n} \\ 0 & \Omega_{32} - \Omega_{31}\Omega_{11}^{-1}\Omega_{12} & \Omega_{33} - \Omega_{31}\Omega_{11}^{-1}\Omega_{13} & \cdots & \Omega_{3n} - \Omega_{31}\Omega_{11}^{-1}\Omega_{1n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & \Omega_{n2} - \Omega_{n1}\Omega_{11}^{-1}\Omega_{12} & \Omega_{n3} - \Omega_{n1}\Omega_{11}^{-1}\Omega_{13} & \cdots & \Omega_{nn} - \Omega_{n1}\Omega_{11}^{-1}\Omega_{1n} \end{bmatrix}.$$

We next proceed in exactly the same way with the second column of $\mathbf{H}$. The approach now will be to multiply the second row of $\mathbf{H}$ by $h_{32}h_{22}^{-1}$ and subtract the result from the third row. Similarly, we multiply the second row of $\mathbf{H}$ by $h_{42}h_{22}^{-1}$ and subtract the result from the fourth row, and so on down through the second

column of $\mathbf{H}$. These operations can be represented as premultiplying $\mathbf{H}$ by the following matrix:

$$\mathbf{E}_2 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & -h_{32}h_{22}^{-1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & -h_{n2}h_{22}^{-1} & 0 & \cdots & 1 \end{bmatrix}. \qquad [4.4.6]$$

This matrix always exists provided that $h_{22} \neq 0$. But h_{22} can be calculated as $h_{22} = \mathbf{e}_2'\mathbf{H}\mathbf{e}_2$, where $\mathbf{e}_2' = [0 \quad 1 \quad 0 \cdots 0]$. Moreover, $\mathbf{H} = \mathbf{E}_1\mathbf{\Omega}\mathbf{E}_1'$, where $\mathbf{\Omega}$ is positive definite and $\mathbf{E}_1$ is given by [4.4.3]. Since $\mathbf{E}_1$ is lower triangular, its determinant is the product of terms along the principal diagonal, which are all unity. Thus $\mathbf{E}_1$ is nonsingular, meaning that $\mathbf{H} = \mathbf{E}_1\mathbf{\Omega}\mathbf{E}_1'$ is positive definite and so $h_{22} = \mathbf{e}_2'\mathbf{H}\mathbf{e}_2$ must be strictly positive. Thus the matrix in [4.4.6] can always be calculated.

If $\mathbf{H}$ is premultiplied by the matrix in [4.4.6] and postmultiplied by the transpose, the result is

$$\mathbf{E}_2\mathbf{H}\mathbf{E}_2' = \mathbf{K},$$

where

$$\mathbf{K} = \begin{bmatrix} h_{11} & 0 & 0 & \cdots & 0 \\ 0 & h_{22} & 0 & \cdots & 0 \\ 0 & 0 & h_{33} - h_{32}h_{22}^{-1}h_{23} & \cdots & h_{3n} - h_{32}h_{22}^{-1}h_{2n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & h_{n3} - h_{n2}h_{22}^{-1}h_{23} & \cdots & h_{nn} - h_{n2}h_{22}^{-1}h_{2n} \end{bmatrix}.$$

Again, since $\mathbf{H}$ is positive definite and since $\mathbf{E}_2$ is nonsingular, $\mathbf{K}$ is positive definite and in particular k_{33} is positive. Proceeding through each of the columns with the same approach, we see that for any positive definite symmetric matrix $\mathbf{\Omega}$ there exist matrices $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_{n-1}$ such that

$$\mathbf{E}_{n-1} \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{\Omega}\mathbf{E}_1'\mathbf{E}_2' \cdots \mathbf{E}_{n-1}' = \mathbf{D}, \qquad [4.4.7]$$

where

$$\mathbf{D} = \begin{bmatrix} \Omega_{11} & 0 & 0 & \cdots & 0 \\ 0 & \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} & 0 & \cdots & 0 \\ 0 & 0 & h_{33} - h_{32}h_{22}^{-1}h_{23} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & c_{nn} - c_{n,n-1}c_{n-1,n-1}^{-1}c_{n-1,n} \end{bmatrix},$$

with all the diagonal entries of $\mathbf{D}$ strictly positive. The matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ in [4.4.7] are given by [4.4.3] and [4.4.6]. In general, $\mathbf{E}_j$ is a matrix with nonzero values in the jth column below the principal diagonal, 1s along the principal diagonal, and zeros everywhere else.

Thus each $\mathbf{E}_j$ is lower triangular with unit determinant. Hence $\mathbf{E}_j^{-1}$ exists, and the following matrix exists:

$$\mathbf{A} = (\mathbf{E}_{n-1} \cdots \mathbf{E}_2\mathbf{E}_1)^{-1} = \mathbf{E}_1^{-1}\mathbf{E}_2^{-1} \cdots \mathbf{E}_{n-1}^{-1}. \qquad [4.4.8]$$

If [4.4.7] is premultiplied by $\mathbf{A}$ and postmultiplied by $\mathbf{A}'$, the result is

$$\boldsymbol{\Omega} = \mathbf{ADA}'. \qquad [4.4.9]$$

Recall that $\mathbf{E}_1$ represents the operation of multiplying the first row of $\boldsymbol{\Omega}$ by certain numbers and subtracting the results from each of the subsequent rows. Its inverse $\mathbf{E}_1^{-1}$ undoes this operation, which would be achieved by multiplying the first row by these same numbers and *adding* the results to the subsequent rows. Thus

$$\mathbf{E}_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \Omega_{21}\Omega_{11}^{-1} & 1 & 0 & \cdots & 0 \\ \Omega_{31}\Omega_{11}^{-1} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \Omega_{n1}\Omega_{11}^{-1} & 0 & 0 & \cdots & 1 \end{bmatrix}, \qquad [4.4.10]$$

as may be verified directly by multiplying [4.4.3] by [4.4.10] to obtain the identity matrix. Similarly,

$$\mathbf{E}_2^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & h_{32}h_{22}^{-1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & h_{n2}h_{22}^{-1} & 0 & \cdots & 1 \end{bmatrix},$$

and so on. Because of this special structure, the series of multiplications in [4.4.8] turns out to be trivial to carry out:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \Omega_{21}\Omega_{11}^{-1} & 1 & 0 & \cdots & 0 \\ \Omega_{31}\Omega_{11}^{-1} & h_{32}h_{22}^{-1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \Omega_{n1}\Omega_{11}^{-1} & h_{n2}h_{22}^{-1} & k_{n3}k_{33}^{-1} & \cdots & 1 \end{bmatrix}. \qquad [4.4.11]$$

That is, the jth column of $\mathbf{A}$ is just the jth column of $\mathbf{E}_j^{-1}$.

We should emphasize that the simplicity of carrying out these matrix multiplications is due not just to the special structure of the $\mathbf{E}_j^{-1}$ matrices but also to the order in which they are multiplied. For example, $\mathbf{A}^{-1} = \mathbf{E}_{n-1}\mathbf{E}_{n-2}\cdots\mathbf{E}_1$ cannot be calculated simply by using the jth column of $\mathbf{E}_j$ for the jth column of $\mathbf{A}^{-1}$.

Since the matrix $\mathbf{A}$ in [4.4.11] is lower triangular with 1s along the principal diagonal, expression [4.4.9] is the triangular factorization of $\boldsymbol{\Omega}$.

For illustration, the triangular factorization $\boldsymbol{\Omega} = \mathbf{ADA}'$ of a (2×2) matrix is

$$\begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \Omega_{21}\Omega_{11}^{-1} & 1 \end{bmatrix} \times \begin{bmatrix} \Omega_{11} & 0 \\ 0 & \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} \end{bmatrix} \begin{bmatrix} 1 & \Omega_{11}^{-1}\Omega_{12} \\ 0 & 1 \end{bmatrix}, \qquad [4.4.12]$$

while that of a (3×3) matrix is

$$\begin{bmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} \\ \Omega_{21} & \Omega_{22} & \Omega_{23} \\ \Omega_{31} & \Omega_{32} & \Omega_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \Omega_{21}\Omega_{11}^{-1} & 1 & 0 \\ \Omega_{31}\Omega_{11}^{-1} & h_{32}h_{22}^{-1} & 1 \end{bmatrix} \times \begin{bmatrix} \Omega_{11} & 0 & 0 \\ 0 & h_{22} & 0 \\ 0 & 0 & h_{33} - h_{32}h_{22}^{-1}h_{23} \end{bmatrix} \begin{bmatrix} 1 & \Omega_{11}^{-1}\Omega_{12} & \Omega_{11}^{-1}\Omega_{13} \\ 0 & 1 & h_{22}^{-1}h_{23} \\ 0 & 0 & 1 \end{bmatrix}, \quad [4.4.13]$$

where $h_{22} = (\Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12})$, $h_{33} = (\Omega_{33} - \Omega_{31}\Omega_{11}^{-1}\Omega_{13})$, and $h_{23} = h_{32} = (\Omega_{23} - \Omega_{21}\Omega_{11}^{-1}\Omega_{13})$.

Uniqueness of the Triangular Factorization

We next establish that the triangular factorization is unique. Suppose that

$$\mathbf{\Omega} = \mathbf{A}_1\mathbf{D}_1\mathbf{A}_1' = \mathbf{A}_2\mathbf{D}_2\mathbf{A}_2', \quad [4.4.14]$$

where $\mathbf{A}_1$ and $\mathbf{A}_2$ are both lower triangular with 1s along the principal diagonal and $\mathbf{D}_1$ and $\mathbf{D}_2$ are both diagonal with positive entries along the principal diagonal. Then all the matrices have inverses. Premultiplying [4.4.14] by $\mathbf{D}_1^{-1}\mathbf{A}_1^{-1}$ and postmultiplying by $[\mathbf{A}_2']^{-1}$ yields

$$\mathbf{A}_1'[\mathbf{A}_2']^{-1} = \mathbf{D}_1^{-1}\mathbf{A}_1^{-1}\mathbf{A}_2\mathbf{D}_2. \quad [4.4.15]$$

Since $\mathbf{A}_2'$ is upper triangular with 1s along the principal diagonal, $[\mathbf{A}_2']^{-1}$ must likewise be upper triangular with 1s along the principal diagonal. Since $\mathbf{A}_1'$ is also of this form, the left side of [4.4.15] is upper triangular with 1s along the principal diagonal. By similar reasoning, the right side of [4.4.15] must be lower triangular. The only way an upper triangular matrix can equal a lower triangular matrix is if all the off-diagonal terms are zero. Moreover, since the diagonal entries on the left side of [4.4.15] are all unity, this matrix must be the identity matrix:

$$\mathbf{A}_1'[\mathbf{A}_2']^{-1} = \mathbf{I}_n.$$

Postmultiplication by $\mathbf{A}_2'$ establishes that $\mathbf{A}_1' = \mathbf{A}_2'$. Premultiplying [4.4.14] by $\mathbf{A}^{-1}$ and postmultiplying by $[\mathbf{A}']^{-1}$ then yields $\mathbf{D}_1 = \mathbf{D}_2$.

The Cholesky Factorization

A closely related factorization of a symmetric positive definite matrix $\mathbf{\Omega}$ is obtained as follows. Define $\mathbf{D}^{1/2}$ to be the $(n \times n)$ diagonal matrix whose diagonal entries are the square roots of the corresponding elements of the matrix $\mathbf{D}$ in the triangular factorization:

$$\mathbf{D}^{1/2} = \begin{bmatrix} \sqrt{d_{11}} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{d_{22}} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{d_{33}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{d_{nn}} \end{bmatrix}.$$

Since the matrix $\mathbf{D}$ is unique and has strictly positive diagonal entries, the matrix $\mathbf{D}^{1/2}$ exists and is unique. Then the triangular factorization can be written

$$\mathbf{\Omega} = \mathbf{A}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{A}' = \mathbf{A}\mathbf{D}^{1/2}(\mathbf{A}\mathbf{D}^{1/2})'$$

or

$$\boldsymbol{\Omega} = \mathbf{PP}', \tag{4.4.16}$$

where

$$\mathbf{P} \equiv \mathbf{AD}^{1/2}$$

$$= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{21} & 1 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sqrt{d_{11}} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{d_{22}} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{d_{33}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{d_{nn}} \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{d_{11}} & 0 & 0 & \cdots & 0 \\ a_{21}\sqrt{d_{11}} & \sqrt{d_{22}} & 0 & \cdots & 0 \\ a_{31}\sqrt{d_{11}} & a_{32}\sqrt{d_{22}} & \sqrt{d_{33}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{n1}\sqrt{d_{11}} & a_{n2}\sqrt{d_{22}} & a_{n3}\sqrt{d_{33}} & \cdots & \sqrt{d_{nn}} \end{bmatrix}.$$

Expression [4.4.16] is known as the *Cholesky factorization* of $\boldsymbol{\Omega}$. Note that $\mathbf{P}$, like $\mathbf{A}$, is lower triangular, though whereas $\mathbf{A}$ has 1s along the principal diagonal, the Cholesky factor has the square roots of the elements of $\mathbf{D}$ along the principal diagonal.

4.5. *Updating a Linear Projection*

Triangular Factorization of a Second-Moment Matrix and Linear Projection

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ be an $(n \times 1)$ vector of random variables whose second-moment matrix is given by

$$\boldsymbol{\Omega} = E(\mathbf{YY}'). \tag{4.5.1}$$

Let $\boldsymbol{\Omega} = \mathbf{ADA}'$ be the triangular factorization of $\boldsymbol{\Omega}$, and define

$$\tilde{\mathbf{Y}} \equiv \mathbf{A}^{-1}\mathbf{Y}. \tag{4.5.2}$$

The second-moment matrix of these transformed variables is given by

$$E(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}') = E(\mathbf{A}^{-1}\mathbf{YY}'[\mathbf{A}']^{-1}) = \mathbf{A}^{-1}E(\mathbf{YY}')[\mathbf{A}']^{-1}. \tag{4.5.3}$$

Substituting [4.5.1] into [4.5.3], the second-moment matrix of $\tilde{\mathbf{Y}}$ is seen to be diagonal:

$$E(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}') = \mathbf{A}^{-1}\boldsymbol{\Omega}[\mathbf{A}']^{-1} = \mathbf{A}^{-1}\mathbf{ADA}'[\mathbf{A}']^{-1} = \mathbf{D}. \tag{4.5.4}$$

That is,

$$E(\tilde{Y}_i\tilde{Y}_j) = \begin{cases} d_{ii} & \text{for } i = j \\ 0 & \text{for } i \neq j. \end{cases} \tag{4.5.5}$$

Thus the $\tilde{Y}$'s form a series of random variables that are uncorrelated with one another.[4] To see the implication of this, premultiply [4.5.2] by $\mathbf{A}$:

$$\mathbf{A}\tilde{\mathbf{Y}} = \mathbf{Y}. \tag{4.5.6}$$

[4]We will use "Y_i and Y_j are uncorrelated" to mean "$E(Y_iY_j) = 0$." The terminology will be correct if Y_i and Y_j have zero means or if a constant term is included in the linear projection.

Expression [4.4.11] can be used to write out [4.5.6] explicitly as

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \Omega_{21}\Omega_{11}^{-1} & 1 & 0 & \cdots & 0 \\ \Omega_{31}\Omega_{11}^{-1} & h_{32}h_{22}^{-1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \Omega_{n1}\Omega_{11}^{-1} & h_{n2}h_{22}^{-1} & k_{n3}k_{33}^{-1} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \\ \tilde{Y}_3 \\ \vdots \\ \tilde{Y}_n \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}. \qquad [4.5.7]$$

The first equation in [4.5.7] states that

$$\tilde{Y}_1 = Y_1, \qquad [4.5.8]$$

so the first elements of the vectors $\mathbf{Y}$ and $\tilde{\mathbf{Y}}$ represent the same random variable.

The second equation in [4.5.7] asserts that

$$\Omega_{21}\Omega_{11}^{-1}\tilde{Y}_1 + \tilde{Y}_2 = Y_2,$$

or, using [4.5.8],

$$\tilde{Y}_2 = Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1 \equiv Y_2 - \alpha Y_1, \qquad [4.5.9]$$

where we have defined $\alpha \equiv \Omega_{21}\Omega_{11}^{-1}$. The fact that $\tilde{Y}_2$ is uncorrelated with $\tilde{Y}_1$ implies

$$E(\tilde{Y}_2\tilde{Y}_1) = E[(Y_2 - \alpha Y_1)Y_1] = 0. \qquad [4.5.10]$$

But, recalling [4.1.10], the value of α that satisfies [4.5.10] is defined as the coefficient of the linear projection of Y_2 on Y_1. Thus the triangular factorization of $\mathbf{\Omega}$ can be used to infer that the coefficient of a linear projection of Y_2 on Y_1 is given by $\alpha = \Omega_{21}\Omega_{11}^{-1}$, confirming the earlier result [4.1.13]. In general, the row i, column 1 entry of $\mathbf{A}$ is $\Omega_{i1}\Omega_{11}^{-1}$, which is the coefficient from a linear projection of Y_i on Y_1.

Since $\tilde{Y}_2$ has the interpretation as the residual from a projection of Y_2 on Y_1, from [4.5.5] d_{22} gives the *MSE* of this projection:

$$E(\tilde{Y}_2^2) = d_{22} = \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}.$$

This confirms the formula for the *MSE* of a linear projection derived earlier (equation [4.1.15]).

The third equation in [4.5.7] states that

$$\Omega_{31}\Omega_{11}^{-1}\tilde{Y}_1 + h_{32}h_{22}^{-1}\tilde{Y}_2 + \tilde{Y}_3 = Y_3.$$

Substituting in from [4.5.8] and [4.5.9] and rearranging,

$$\tilde{Y}_3 = Y_3 - \Omega_{31}\Omega_{11}^{-1}Y_1 - h_{32}h_{22}^{-1}(Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1). \qquad [4.5.11]$$

Thus $\tilde{Y}_3$ is the residual from subtracting a particular linear combination of Y_1 and Y_2 from Y_3. From [4.5.5], this residual is uncorrelated with either $\tilde{Y}_1$ or $\tilde{Y}_2$:

$$E[Y_3 - \Omega_{31}\Omega_{11}^{-1}Y_1 - h_{32}h_{22}^{-1}(Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1)]\tilde{Y}_j = 0 \qquad \text{for } j = 1 \text{ or } 2.$$

Thus this residual is uncorrelated with either Y_1 or Y_2, meaning that $\tilde{Y}_3$ has the interpretation as the residual from a linear projection of Y_3 on Y_1 and Y_2. According to [4.5.11], the linear projection is given by

$$\hat{P}(Y_3|Y_2,Y_1) = \Omega_{31}\Omega_{11}^{-1}Y_1 + h_{32}h_{22}^{-1}(Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1). \qquad [4.5.12]$$

The *MSE* of the linear projection is the variance of $\tilde{Y}_3$, which from [4.5.5] is given by d_{33}:

$$E[Y_3 - \hat{P}(Y_3|Y_2,Y_1)]^2 = h_{33} - h_{32}h_{22}^{-1}h_{23}. \qquad [4.5.13]$$

Expression [4.5.12] gives a convenient formula for updating a linear projection. Suppose we are interested in forecasting the value of Y_3. Let Y_1 be some initial information on which this forecast might be based. A forecast of Y_3 on the basis of Y_1 alone takes the form

$$\hat{P}(Y_3|Y_1) = \Omega_{31}\Omega_{11}^{-1}Y_1.$$

Let Y_2 represent some new information with which we could update this forecast. If we were asked to guess the magnitude of this second variable on the basis of Y_1 alone, the answer would be

$$\hat{P}(Y_2|Y_1) = \Omega_{21}\Omega_{11}^{-1}Y_1.$$

Equation [4.5.12] states that

$$\hat{P}(Y_3|Y_2,Y_1) = \hat{P}(Y_3|Y_1) + h_{32}h_{22}^{-1}[Y_2 - \hat{P}(Y_2|Y_1)]. \qquad [4.5.14]$$

We can thus optimally update the initial forecast $\hat{P}(Y_3|Y_1)$ by adding to it a multiple $(h_{32}h_{22}^{-1})$ of the unanticipated component of the new information $[Y_2 - \hat{P}(Y_2|Y_1)]$. This multiple $(h_{32}h_{22}^{-1})$ can also be interpreted as the coefficient on Y_2 in a linear projection of Y_3 on Y_2 and Y_1.

To understand the nature of the multiplier $(h_{32}h_{22}^{-1})$, define the $(n \times 1)$ vector $\tilde{\mathbf{Y}}(1)$ by

$$\tilde{\mathbf{Y}}(1) \equiv \mathbf{E}_1\mathbf{Y}, \qquad [4.5.15]$$

where $\mathbf{E}_1$ is the matrix given in [4.4.3]. Notice that the second-moment matrix of $\tilde{\mathbf{Y}}(1)$ is given by

$$E\{\tilde{\mathbf{Y}}(1)[\tilde{\mathbf{Y}}(1)]'\} = E\{\mathbf{E}_1\mathbf{Y}\mathbf{Y}'\mathbf{E}_1'\} = \mathbf{E}_1\mathbf{\Omega}\mathbf{E}_1'.$$

But from [4.4.4] this is just the matrix $\mathbf{H}$. Thus $\mathbf{H}$ has the interpretation as the second-moment matrix of $\tilde{\mathbf{Y}}(1)$. Substituting [4.4.3] into [4.5.15],

$$\tilde{\mathbf{Y}}(1) = \begin{bmatrix} Y_1 \\ Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1 \\ Y_3 - \Omega_{31}\Omega_{11}^{-1}Y_1 \\ \vdots \\ Y_n - \Omega_{n1}\Omega_{11}^{-1}Y_1 \end{bmatrix}.$$

The first element of $\tilde{\mathbf{Y}}(1)$ is thus just Y_1 itself, while the ith element of $\tilde{\mathbf{Y}}(1)$ for $i = 2, 3, \ldots, n$ is the residual from a projection of Y_i on Y_1. The matrix $\mathbf{H}$ is thus the second-moment matrix of the residuals from projections of each of the variables on Y_1. In particular, h_{22} is the *MSE* from a projection of Y_2 on Y_1:

$$h_{22} = E[Y_2 - \hat{P}(Y_2|Y_1)]^2,$$

while h_{32} is the expected product of this error with the error from a projection of Y_3 on Y_1:

$$h_{32} = E\{[Y_3 - \hat{P}(Y_3|Y_1)][Y_2 - \hat{P}(Y_2|Y_1)]\}.$$

Thus equation [4.5.14] states that a linear projection can be updated using the following formula:

$$\begin{aligned}\hat{P}(Y_3|Y_2,Y_1) = {} & \hat{P}(Y_3|Y_1) \\ & + \{E[Y_3 - \hat{P}(Y_3|Y_1)][Y_2 - \hat{P}(Y_2|Y_1)]\} \\ & \times \{E[Y_2 - \hat{P}(Y_2|Y_1)]^2\}^{-1} \times [Y_2 - \hat{P}(Y_2|Y_1)].\end{aligned} \qquad [4.5.16]$$

For example, suppose that Y_1 is a constant term, so that $\hat{P}(Y_2|Y_1)$ is just μ_2, the mean of Y_2, while $\hat{P}(Y_3|Y_1) = \mu_3$. Equation [4.5.16] then states that

$$\hat{P}(Y_3|Y_2,1) = \mu_3 + \text{Cov}(Y_3, Y_2)\cdot[\text{Var}(Y_2)]^{-1}\cdot(Y_2 - \mu_2).$$

The *MSE* associated with this updated linear projection can also be calculated from the triangular factorization. From [4.5.5], the *MSE* from a linear projection of Y_3 on Y_2 and Y_1 can be calculated from

$$\begin{aligned} E[Y_3 - \hat{P}(Y_3|Y_2,Y_1)]^2 &= E(\hat{Y}_3^2) \\ &= d_{33} \\ &= h_{33} - h_{32}h_{22}^{-1}h_{23}. \end{aligned}$$

In general, for $i > 2$, the coefficient on Y_2 in a linear projection of Y_i on Y_2 and Y_1 is given by the ith element of the second column of the matrix **A**. For any $i > j$, the coefficients on Y_j in a linear projection of Y_i on Y_j, Y_{j-1}, . . . , Y_1 is given by the row i, column j element of **A**. The magnitude d_{ii} gives the *MSE* for a linear projection of Y_i on Y_{i-1}, Y_{i-2}, . . . , Y_1.

Application: Exact Finite-Sample Forecasts for an MA*(1) Process*

As an example of applying these results, suppose that Y_t follows an $MA(1)$ process:

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1},$$

where ε_t is a white noise process with variance σ^2 and θ is unrestricted. Suppose we want to forecast the value of Y_n on the basis of the previous $n - 1$ values (Y_1, Y_2, . . . , Y_{n-1}). Let

$$\mathbf{Y}' \equiv [(Y_1 - \mu) \quad (Y_2 - \mu) \quad \cdots \quad (Y_{n-1} - \mu) \quad (Y_n - \mu)],$$

and let $\boldsymbol{\Omega}$ denote the $(n \times n)$ variance-covariance matrix of **Y**:

$$\boldsymbol{\Omega} = E(\mathbf{Y}\mathbf{Y}') = \sigma^2\begin{bmatrix} 1+\theta^2 & \theta & 0 & \cdots & 0 \\ \theta & 1+\theta^2 & \theta & \cdots & 0 \\ 0 & \theta & 1+\theta^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\theta^2 \end{bmatrix}. \qquad [4.5.17]$$

Appendix 4.B to this chapter shows that the triangular factorization of $\boldsymbol{\Omega}$ is

$$\mathbf{A} = \qquad [4.5.18]$$

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ \dfrac{\theta}{1+\theta^2} & 1 & 0 & \cdots & 0 & 0 \\ 0 & \dfrac{\theta(1+\theta^2)}{1+\theta^2+\theta^4} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \dfrac{\theta[1+\theta^2+\theta^4+\cdots+\theta^{2(n-2)}]}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}} & 1 \end{bmatrix}$$

$$\mathbf{D} = \qquad [4.5.19]$$

$$\sigma^2 \begin{bmatrix} 1+\theta^2 & 0 & 0 & \cdots & 0 \\ 0 & \dfrac{1+\theta^2+\theta^4}{1+\theta^2} & 0 & \cdots & 0 \\ 0 & 0 & \dfrac{1+\theta^2+\theta^4+\theta^6}{1+\theta^2+\theta^4} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \dfrac{1+\theta^2+\theta^4+\cdots+\theta^{2n}}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}} \end{bmatrix}.$$

To use the triangular factorization to calculate exact finite-sample forecasts, recall that $\tilde{Y}_i$, the ith element of $\tilde{\mathbf{Y}} = \mathbf{A}^{-1}\mathbf{Y}$, has the interpretation as the residual from a linear projection of Y_i on a constant and its previous values:

$$\tilde{Y}_i = Y_i - \hat{E}(Y_i|Y_{i-1}, Y_{i-2}, \ldots, Y_1).$$

The system of equations $\mathbf{A}\tilde{\mathbf{Y}} = \mathbf{Y}$ can be written out explicitly as

$$\begin{aligned} \tilde{Y}_1 &= Y_1 - \mu \\ \frac{\theta}{1+\theta^2}\tilde{Y}_1 + \tilde{Y}_2 &= Y_2 - \mu \\ \frac{\theta(1+\theta^2)}{1+\theta^2+\theta^4}\tilde{Y}_2 + \tilde{Y}_3 &= Y_3 - \mu \\ &\vdots \\ \frac{\theta[1+\theta^2+\theta^4+\cdots+\theta^{2(n-2)}]}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}}\tilde{Y}_{n-1} + \tilde{Y}_n &= Y_n - \mu. \end{aligned}$$

Solving the last equation for $\tilde{Y}_n$,

$$\begin{aligned} Y_n - \hat{E}(Y_n|Y_{n-1}, Y_{n-2}, \ldots, Y_1) &= Y_n - \mu \\ - \frac{\theta[1+\theta^2+\theta^4+\cdots+\theta^{2(n-2)}]}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}} &[Y_{n-1} - \hat{E}(Y_{n-1}|Y_{n-2}, Y_{n-3}, \ldots, Y_1)], \end{aligned}$$

implying

$$\begin{aligned} \hat{E}(Y_n|Y_{n-1}, Y_{n-2}, \ldots, Y_1) &= \mu \qquad [4.5.20] \\ + \frac{\theta[1+\theta^2+\theta^4+\cdots+\theta^{2(n-2)}]}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}} &[Y_{n-1} - \hat{E}(Y_{n-1}|Y_{n-2}, Y_{n-3}, \ldots, Y_1)]. \end{aligned}$$

The *MSE* of this forecast is given by d_{nn}:

$$MSE[\hat{E}(Y_n|Y_{n-1}, Y_{n-2}, \ldots, Y_1)] = \sigma^2 \frac{1+\theta^2+\theta^4+\cdots+\theta^{2n}}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}}. \qquad [4.5.21]$$

It is interesting to note the behavior of this optimal forecast as the number of observations (n) becomes large. First, suppose that the moving average representation is invertible ($|\theta| < 1$). In this case, as $n \to \infty$, the coefficient in [4.5.20] tends to θ:

$$\frac{\theta[1+\theta^2+\theta^4+\cdots+\theta^{2(n-2)}]}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}} \to \theta,$$

while the *MSE* [4.5.21] tends to σ^2, the variance of the fundamental innovation. Thus the optimal forecast for a finite number of observations [4.5.20] eventually tends toward the forecast rule used for an infinite number of observations [4.2.32].

Alternatively, the calculations that produced [4.5.20] are equally valid for a noninvertible representation with $|\theta| > 1$. In this case the coefficient in [4.5.20] tends toward θ^{-1}:

$$\begin{aligned}\frac{\theta[1 + \theta^2 + \theta^4 + \cdots + \theta^{2(n-2)}]}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2(n-1)}} &= \frac{\theta[1 - \theta^{2(n-1)}]/(1 - \theta^2)}{(1 - \theta^{2n})/(1 - \theta^2)} \\ &= \frac{\theta(\theta^{-2n} - \theta^{-2})}{\theta^{-2n} - 1} \\ &\to \frac{\theta(-\theta^{-2})}{-1} \\ &= \theta^{-1}.\end{aligned}$$

Thus, the coefficient in [4.5.20] tends to θ^{-1} in this case, which is the moving average coefficient associated with the invertible representation. The *MSE* [4.5.21] tends to $\sigma^2\theta^2$:

$$\sigma^2 \frac{[1 - \theta^{2(n+1)}]/(1 - \theta^2)}{(1 - \theta^{2n})/(1 - \theta^2)} \to \sigma^2\theta^2,$$

which will be recognized from [3.7.7] as the variance of the innovation associated with the fundamental representation.

This observation explains the use of the expression "fundamental" in this context. The fundamental innovation ε_t has the property that

$$Y_t - \hat{E}(Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_{t-m}) \xrightarrow{m.s.} \varepsilon_t \qquad [4.5.22]$$

as $m \to \infty$ where $\xrightarrow{m.s.}$ denotes mean square convergence. Thus when $|\theta| > 1$, the coefficient θ in the approximation in [4.3.3] should be replaced by θ^{-1}. When this is done, expression [4.3.3] will approach the correct forecast as $m \to \infty$.

It is also instructive to consider the borderline case $\theta = 1$. The optimal finite-sample forecast for an $MA(1)$ process with $\theta = 1$ is seen from [4.5.20] to be given by

$$\hat{E}(Y_n|Y_{n-1}, Y_{n-2}, \ldots, Y_1) = \mu + \frac{n-1}{n}[Y_{n-1} - \hat{E}(Y_{n-1}|Y_{n-2}, Y_{n-3}, \ldots, Y_1)],$$

which, after recursive substitution, becomes

$$\begin{aligned}&\hat{E}(Y_n|Y_{n-1}, Y_{n-2}, \ldots, Y_1) \\ &\quad = \mu + \frac{n-1}{n}(Y_{n-1} - \mu) - \frac{n-2}{n}(Y_{n-2} - \mu) \\ &\qquad + \frac{n-3}{n}(Y_{n-3} - \mu) - \cdots + (-1)^n \frac{1}{n}(Y_1 - \mu).\end{aligned} \qquad [4.5.23]$$

The *MSE* of this forecast is given by [4.5.21]:

$$\sigma^2(n + 1)/n \to \sigma^2.$$

Thus the variance of the forecast error again tends toward that of ε_t. Hence the innovation ε_t is again fundamental for this case in the sense of [4.5.22]. Note the contrast between the optimal forecast [4.5.23] and a forecast based on a naive application of [4.3.3],

$$\mu + (Y_{n-1} - \mu) - (Y_{n-2} - \mu) + (Y_{n-3} - \mu) - \cdots + (-1)^n(Y_1 - \mu). \qquad [4.5.24]$$

The approximation [4.3.3] was derived under the assumption that the moving average representation was invertible, and the borderline case $\theta = 1$ is not invertible. For this

reason [4.5.24] does not converge to the optimal forecast [4.5.23] as n grows large. When $\theta = 1$, $Y_t = \mu + \varepsilon_t + \varepsilon_{t-1}$ and [4.5.24] can be written as

$$\mu + (\varepsilon_{n-1} + \varepsilon_{n-2}) - (\varepsilon_{n-2} + \varepsilon_{n-3}) + (\varepsilon_{n-3} + \varepsilon_{n-4}) - \cdots + (-1)^n(\varepsilon_1 + \varepsilon_0) = \mu + \varepsilon_{n-1} + (-1)^n\varepsilon_0.$$

The difference between this and Y_n, the value being forecast, is $\varepsilon_n - (-1)^n\varepsilon_0$, which has *MSE* $2\sigma^2$ for all n. Thus, whereas [4.5.23] converges to the optimal forecast as $n \to \infty$, [4.5.24] does not.

Block Triangular Factorization

Suppose we have observations on two sets of variables. The first set of variables is collected in an $(n_1 \times 1)$ vector $\mathbf{Y}_1$ and the second set in an $(n_2 \times 1)$ vector $\mathbf{Y}_2$. Their second-moment matrix can be written in partitioned form as

$$\boldsymbol{\Omega} \equiv \begin{bmatrix} E(\mathbf{Y}_1\mathbf{Y}_1') & E(\mathbf{Y}_1\mathbf{Y}_2') \\ E(\mathbf{Y}_2\mathbf{Y}_1') & E(\mathbf{Y}_2\mathbf{Y}_2') \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix},$$

where $\boldsymbol{\Omega}_{11}$ is an $(n_1 \times n_1)$ matrix, $\boldsymbol{\Omega}_{22}$ is an $(n_2 \times n_2)$ matrix, and the $(n_1 \times n_2)$ matrix $\boldsymbol{\Omega}_{12}$ is the transpose of the $(n_2 \times n_1)$ matrix $\boldsymbol{\Omega}_{21}$.

We can put zeros in the lower left $(n_2 \times n_1)$ block of $\boldsymbol{\Omega}$ by premultiplying $\boldsymbol{\Omega}$ by the following matrix:

$$\overline{\mathbf{E}}_1 = \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ -\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} \end{bmatrix}.$$

If $\boldsymbol{\Omega}$ is premultiplied by $\overline{\mathbf{E}}_1$ and postmultiplied by $\overline{\mathbf{E}}_1'$, the result is

$$\begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ -\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{n_1} & -\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} \\ \mathbf{0} & \mathbf{I}_{n_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} \end{bmatrix}. \quad [4.5.25]$$

Define

$$\overline{\mathbf{A}} \equiv \overline{\mathbf{E}}_1^{-1} = \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} \end{bmatrix}.$$

If [4.5.25] is premultiplied by $\overline{\mathbf{A}}$ and postmultiplied by $\overline{\mathbf{A}}'$, the result is

$$\begin{aligned} \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix} &= \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} \end{bmatrix} \\ &\quad \times \begin{bmatrix} \boldsymbol{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{n_1} & \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} \\ \mathbf{0} & \mathbf{I}_{n_2} \end{bmatrix} \quad [4.5.26] \\ &= \overline{\mathbf{A}}\,\overline{\mathbf{D}}\,\overline{\mathbf{A}}'. \end{aligned}$$

This is similar to the triangular factorization $\boldsymbol{\Omega} = \mathbf{A}\mathbf{D}\mathbf{A}'$, except that $\overline{\mathbf{D}}$ is a block-diagonal matrix rather than a truly diagonal matrix:

$$\overline{\mathbf{D}} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} \end{bmatrix}.$$

As in the earlier case, $\bar{\mathbf{D}}$ can be interpreted as the second-moment matrix of the vector $\tilde{\mathbf{Y}} = \bar{\mathbf{A}}^{-1}\mathbf{Y}$,

$$\begin{bmatrix} \tilde{\mathbf{Y}}_1 \\ \tilde{\mathbf{Y}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ -\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} \end{bmatrix} \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix};$$

that is, $\tilde{\mathbf{Y}}_1 = \mathbf{Y}_1$ and $\tilde{\mathbf{Y}}_2 = \mathbf{Y}_2 - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\mathbf{Y}_1$. The ith element of $\tilde{\mathbf{Y}}_2$ is given by Y_{2i} minus a linear combination of the elements of $\mathbf{Y}_1$. The block-diagonality of $\bar{\mathbf{D}}$ implies that the product of any element of $\tilde{\mathbf{Y}}_2$ with any element of $\mathbf{Y}_1$ has expectation zero. Thus $\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}$ gives the matrix of coefficients associated with the linear projection of the vector $\mathbf{Y}_2$ on the vector $\mathbf{Y}_1$,

$$\hat{P}(\mathbf{Y}_2|\mathbf{Y}_1) = \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\mathbf{Y}_1, \tag{4.5.27}$$

as claimed in [4.1.23]. The *MSE* matrix associated with this linear projection is

$$\begin{aligned} E\{[\mathbf{Y}_2 - \hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)][\mathbf{Y}_2 - \hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)]'\} &= E(\tilde{\mathbf{Y}}_2\tilde{\mathbf{Y}}_2') \\ &= \bar{\mathbf{D}}_{22} \\ &= \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}, \end{aligned} \tag{4.5.28}$$

as claimed in [4.1.24].

The calculations for a (3 × 3) matrix similarly extend to a (3 × 3) block matrix without complications. Let $\mathbf{Y}_1$, $\mathbf{Y}_2$, and $\mathbf{Y}_3$ be $(n_1 \times 1)$, $(n_2 \times 1)$, and $(n_3 \times 1)$ vectors. A block-triangular factorization of their second-moment matrix is obtained from a simple generalization of equation [4.4.13]:

$$\begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} & \boldsymbol{\Omega}_{13} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} & \boldsymbol{\Omega}_{23} \\ \boldsymbol{\Omega}_{31} & \boldsymbol{\Omega}_{32} & \boldsymbol{\Omega}_{33} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} & \mathbf{0} \\ \boldsymbol{\Omega}_{31}\boldsymbol{\Omega}_{11}^{-1} & \mathbf{H}_{32}\mathbf{H}_{22}^{-1} & \mathbf{I}_{n_3} \end{bmatrix}$$

$$\times \begin{bmatrix} \boldsymbol{\Omega}_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{H}_{33} - \mathbf{H}_{32}\mathbf{H}_{22}^{-1}\mathbf{H}_{23} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{n_1} & \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} & \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{13} \\ \mathbf{0} & \mathbf{I}_{n_2} & \mathbf{H}_{22}^{-1}\mathbf{H}_{23} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{n_3} \end{bmatrix} \tag{4.5.29}$$

where $\mathbf{H}_{22} = (\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})$, $\mathbf{H}_{33} = (\boldsymbol{\Omega}_{33} - \boldsymbol{\Omega}_{31}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{13})$, and $\mathbf{H}_{23} = \mathbf{H}_{32}'$ $= (\boldsymbol{\Omega}_{23} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{13})$.

This allows us to generalize the earlier result [4.5.12] on updating a linear projection. The optimal forecast of $\mathbf{Y}_3$ conditional on $\mathbf{Y}_2$ and $\mathbf{Y}_1$ can be read off the last block row of $\bar{\mathbf{A}}$:

$$\begin{aligned} \hat{P}(\mathbf{Y}_3|\mathbf{Y}_2,\mathbf{Y}_1) &= \boldsymbol{\Omega}_{31}\boldsymbol{\Omega}_{11}^{-1}\mathbf{Y}_1 + \mathbf{H}_{32}\mathbf{H}_{22}^{-1}(\mathbf{Y}_2 - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\mathbf{Y}_1) \\ &= \hat{P}(\mathbf{Y}_3|\mathbf{Y}_1) + \mathbf{H}_{32}\mathbf{H}_{22}^{-1}[\mathbf{Y}_2 - \hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)], \end{aligned} \tag{4.5.30}$$

where

$$\begin{aligned} \mathbf{H}_{22} &= E\{[\mathbf{Y}_2 - \hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)][\mathbf{Y}_2 - \hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)]'\} \\ \mathbf{H}_{32} &= E\{[\mathbf{Y}_3 - \hat{P}(\mathbf{Y}_3|\mathbf{Y}_1)][\mathbf{Y}_2 - \hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)]'\}. \end{aligned}$$

The *MSE* of this forecast is the matrix generalization of [4.5.13],

$$E\{[\mathbf{Y}_3 - \hat{P}(\mathbf{Y}_3|\mathbf{Y}_2,\mathbf{Y}_1)][\mathbf{Y}_3 - \hat{P}(\mathbf{Y}_3|\mathbf{Y}_2,\mathbf{Y}_1)]'\} = \mathbf{H}_{33} - \mathbf{H}_{32}\mathbf{H}_{22}^{-1}\mathbf{H}_{23}, \tag{4.5.31}$$

where

$$\mathbf{H}_{33} = E\{[\mathbf{Y}_3 - \hat{P}(\mathbf{Y}_3|\mathbf{Y}_1)][\mathbf{Y}_3 - \hat{P}(\mathbf{Y}_3|\mathbf{Y}_1)]'\}.$$

Law of Iterated Projections

Another useful result, the law of iterated projections, can be inferred immediately from [4.5.30]. What happens if the projection $\hat{P}(\mathbf{Y}_3|\mathbf{Y}_2,\mathbf{Y}_1)$ is itself projected on $\mathbf{Y}_1$? The *law of iterated projections* says that this projection is equal to the simple projection of $\mathbf{Y}_3$ on $\mathbf{Y}_1$:

$$\hat{P}[\hat{P}(\mathbf{Y}_3|\mathbf{Y}_2,\mathbf{Y}_1)|\mathbf{Y}_1] = \hat{P}(\mathbf{Y}_3|\mathbf{Y}_1). \qquad [4.5.32]$$

To verify this claim, we need to show that the difference between $\hat{P}(\mathbf{Y}_3|\mathbf{Y}_2,\mathbf{Y}_1)$ and $\hat{P}(\mathbf{Y}_3|\mathbf{Y}_1)$ is uncorrelated with $\mathbf{Y}_1$. But from [4.5.30], this difference is given by

$$\hat{P}(\mathbf{Y}_3|\mathbf{Y}_2,\mathbf{Y}_1) - \hat{P}(\mathbf{Y}_3|\mathbf{Y}_1) = \mathbf{H}_{32}\mathbf{H}_{22}^{-1}[\mathbf{Y}_2 - \hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)],$$

which indeed is uncorrelated with $\mathbf{Y}_1$ by the definition of the linear projection $\hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)$.

4.6. *Optimal Forecasts for Gaussian Processes*

The forecasting rules developed in this chapter are optimal within the class of linear functions of the variables on which the forecast is based. For Gaussian processes, we can make the stronger claim that as long as a constant term is included among the variables on which the forecast is based, the optimal unrestricted forecast turns out to have a linear form and thus is given by the linear projection.

To verify this, let $\mathbf{Y}_1$ be an $(n_1 \times 1)$ vector with mean $\boldsymbol{\mu}_1$, and $\mathbf{Y}_2$ an $(n_2 \times 1)$ vector with mean $\boldsymbol{\mu}_2$, where the variance-covariance matrix is given by

$$\begin{bmatrix} E(\mathbf{Y}_1 - \boldsymbol{\mu}_1)(\mathbf{Y}_1 - \boldsymbol{\mu}_1)' & E(\mathbf{Y}_1 - \boldsymbol{\mu}_1)(\mathbf{Y}_2 - \boldsymbol{\mu}_2)' \\ E(\mathbf{Y}_2 - \boldsymbol{\mu}_2)(\mathbf{Y}_1 - \boldsymbol{\mu}_1)' & E(\mathbf{Y}_2 - \boldsymbol{\mu}_2)(\mathbf{Y}_2 - \boldsymbol{\mu}_2)' \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix}.$$

If $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are Gaussian, then the joint probability density is

$$f_{\mathbf{Y}_1,\mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2) = \frac{1}{(2\pi)^{(n_1+n_2)/2}} \begin{vmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{vmatrix}^{-1/2} \qquad [4.6.1]$$
$$\times \exp\left\{-\frac{1}{2}[(\mathbf{y}_1 - \boldsymbol{\mu}_1)'\ (\mathbf{y}_2 - \boldsymbol{\mu}_2)'] \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y}_1 - \boldsymbol{\mu}_1 \\ \mathbf{y}_2 - \boldsymbol{\mu}_2 \end{bmatrix}\right\}.$$

The inverse of $\boldsymbol{\Omega}$ is readily found by inverting [4.5.26]:

$$\begin{aligned} \boldsymbol{\Omega}^{-1} &= [\overline{\mathbf{A}}\,\overline{\mathbf{D}}\,\overline{\mathbf{A}}']^{-1} \\ &= [\overline{\mathbf{A}}']^{-1}\overline{\mathbf{D}}^{-1}\overline{\mathbf{A}}^{-1} \\ &= \begin{bmatrix} \mathbf{I}_{n_1} & -\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} \\ \mathbf{0} & \mathbf{I}_{n_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Omega}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1} \end{bmatrix} \qquad [4.6.2] \\ &\quad \times \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ -\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} \end{bmatrix}. \end{aligned}$$

Likewise, the determinant of $\boldsymbol{\Omega}$ can be found by taking the determinant of [4.5.26]:

$$|\boldsymbol{\Omega}| = |\overline{\mathbf{A}}| \cdot |\overline{\mathbf{D}}| \cdot |\overline{\mathbf{A}}'|.$$

But $\bar{\mathbf{A}}$ is a lower triangular matrix. Its determinant is therefore given by the product of terms along the principal diagonal, all of which are unity. Hence $|\bar{\mathbf{A}}| = 1$ and $|\mathbf{\Omega}| = |\bar{\mathbf{D}}|$:[5]

$$\begin{vmatrix} \mathbf{\Omega}_{11} & \mathbf{\Omega}_{12} \\ \mathbf{\Omega}_{21} & \mathbf{\Omega}_{22} \end{vmatrix} = \begin{vmatrix} \mathbf{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12} \end{vmatrix} \qquad [4.6.3]$$
$$= |\mathbf{\Omega}_{11}| \cdot |\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}|.$$

Substituting [4.6.2] and [4.6.3] into [4.6.1], the joint density can be written

$$\begin{aligned} & f_{\mathbf{Y}_1,\mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2) \\ &= \frac{1}{(2\pi)^{(n_1+n_2)/2}} |\mathbf{\Omega}_{11}|^{-1/2} \cdot |\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}|^{-1/2} \\ &\quad \times \exp\left\{ -\frac{1}{2} [(\mathbf{y}_1 - \boldsymbol{\mu}_1)' (\mathbf{y}_2 - \boldsymbol{\mu}_2)'] \begin{bmatrix} \mathbf{I}_{n_1} & -\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12} \\ \mathbf{0} & \mathbf{I}_{n_2} \end{bmatrix} \right. \\ &\quad \left. \times \begin{bmatrix} \mathbf{\Omega}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ -\mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 - \boldsymbol{\mu}_1 \\ \mathbf{y}_2 - \boldsymbol{\mu}_2 \end{bmatrix} \right\} \\ &= \frac{1}{(2\pi)^{(n_1+n_2)/2}} |\mathbf{\Omega}_{11}|^{-1/2} \cdot |\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}|^{-1/2} \\ &\quad \times \exp\left\{ -\frac{1}{2} [(\mathbf{y}_1 - \boldsymbol{\mu}_1)' (\mathbf{y}_2 - \mathbf{m})'] \right. \qquad [4.6.4] \\ &\quad \left. \times \begin{bmatrix} \mathbf{\Omega}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 - \boldsymbol{\mu}_1 \\ \mathbf{y}_2 - \mathbf{m} \end{bmatrix} \right\} \\ &= \frac{1}{(2\pi)^{(n_1+n_2)/2}} |\mathbf{\Omega}_{11}|^{-1/2} \cdot |\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}|^{-1/2} \\ &\quad \times \exp\left\{ -\frac{1}{2} (\mathbf{y}_1 - \boldsymbol{\mu}_1)'\mathbf{\Omega}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1) \right. \\ &\quad \left. - \frac{1}{2} (\mathbf{y}_2 - \mathbf{m})'(\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12})^{-1}(\mathbf{y}_2 - \mathbf{m}) \right\}, \end{aligned}$$

where

$$\mathbf{m} \equiv \boldsymbol{\mu}_2 + \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1). \qquad [4.6.5]$$

The conditional density of $\mathbf{Y}_2$ given $\mathbf{Y}_1$ is found by dividing the joint density [4.6.4] by the marginal density:

$$f_{\mathbf{Y}_1}(\mathbf{y}_1) = \frac{1}{(2\pi)^{n_1/2}} |\mathbf{\Omega}_{11}|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{y}_1 - \boldsymbol{\mu}_1)'\mathbf{\Omega}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1) \right].$$

[5]Write $\mathbf{\Omega}_{11}$ in Jordan form as $\mathbf{M}_1\mathbf{J}_1\mathbf{M}_1^{-1}$, where $\mathbf{J}_1$ is upper triangular with eigenvalues of $\mathbf{\Omega}_{11}$ along the principal diagonal. Write $\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}$ as $\mathbf{M}_2\mathbf{J}_2\mathbf{M}_2^{-1}$. Then $\mathbf{\Omega} = \mathbf{M}\mathbf{J}\mathbf{M}^{-1}$, where

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_2 \end{bmatrix} \qquad \mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{bmatrix}.$$

Thus $\mathbf{\Omega}$ has the same determinant as $\mathbf{J}$. Because $\mathbf{J}$ is upper triangular, its determinant is the product of terms along the principal diagonal, or $|\mathbf{J}| = |\mathbf{J}_1| \cdot |\mathbf{J}_2|$. Hence $|\mathbf{\Omega}| = |\mathbf{\Omega}_{11}| \cdot |\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}|$.

The result of this division is

$$f_{\mathbf{Y}_2|\mathbf{Y}_1}(\mathbf{y}_2|\mathbf{y}_1) = \frac{f_{\mathbf{Y}_1,\mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2)}{f_{\mathbf{Y}_1}(\mathbf{y}_1)}$$

$$= \frac{1}{(2\pi)^{n_2/2}} |\mathbf{H}|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{y}_2 - \mathbf{m})'\mathbf{H}^{-1}(\mathbf{y}_2 - \mathbf{m}) \right],$$

where

$$\mathbf{H} \equiv \mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}. \qquad [4.6.6]$$

In other words,

$$\mathbf{Y}_2|\mathbf{Y}_1 \sim N(\mathbf{m}, \mathbf{H})$$

$$\sim N\Big([\boldsymbol{\mu}_2 + \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1)], [\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}] \Big). \qquad [4.6.7]$$

We saw in Section 4.1 that the optimal unrestricted forecast is given by the conditional expectation. For a Gaussian process, the optimal forecast is thus

$$E(\mathbf{Y}_2|\mathbf{Y}_1) = \boldsymbol{\mu}_2 + \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1).$$

On the other hand, for any distribution, the linear projection of the vector $\mathbf{Y}_2$ on a vector $\mathbf{Y}_1$ and a constant term is given by

$$\hat{E}(\mathbf{Y}_2|\mathbf{Y}_1) = \boldsymbol{\mu}_2 + \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1).$$

Hence, for a Gaussian process, the linear projection gives the unrestricted optimal forecast.

4.7. *Sums of* ARMA *Processes*

This section explores the nature of series that result from adding two different *ARMA* processes together, beginning with an instructive example.

Sum of an MA*(1) Process Plus White Noise*

Suppose that a series X_t follows a zero-mean $MA(1)$ process:

$$X_t = u_t + \delta u_{t-1}, \qquad [4.7.1]$$

where u_t is white noise:

$$E(u_t u_{t-j}) = \begin{cases} \sigma_u^2 & \text{for } j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

The autocovariances of X_t are thus

$$E(X_t X_{t-j}) = \begin{cases} (1 + \delta^2)\sigma_u^2 & \text{for } j = 0 \\ \delta\sigma_u^2 & \text{for } j = \pm 1 \\ 0 & \text{otherwise.} \end{cases} \qquad [4.7.2]$$

Let v_t indicate a separate white noise series:

$$E(v_t v_{t-j}) = \begin{cases} \sigma_v^2 & \text{for } j = 0 \\ 0 & \text{otherwise.} \end{cases} \qquad [4.7.3]$$

Suppose, furthermore, that v and u are uncorrelated at all leads and lags:

$$E(u_t v_{t-j}) = 0 \qquad \text{for all } j,$$

implying

$$E(X_t v_{t-j}) = 0 \qquad \text{for all } j. \tag{4.7.4}$$

Let an observed series Y_t represent the sum of the $MA(1)$ and the white noise process:

$$\begin{aligned} Y_t &= X_t + v_t \\ &= u_t + \delta u_{t-1} + v_t. \end{aligned} \tag{4.7.5}$$

The question now posed is, What are the time series properties of Y?

Clearly, Y_t has mean zero, and its autocovariances can be deduced from [4.7.2] through [4.7.4]:

$$\begin{aligned} E(Y_t Y_{t-j}) &= E(X_t + v_t)(X_{t-j} + v_{t-j}) \\ &= E(X_t X_{t-j}) + E(v_t v_{t-j}) \\ &= \begin{cases} (1 + \delta^2)\sigma_u^2 + \sigma_v^2 & \text{for } j = 0 \\ \delta\sigma_u^2 & \text{for } j = \pm 1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{4.7.6}$$

Thus, the sum $X_t + v_t$ is covariance-stationary, and its autocovariances are zero beyond one lag, as are those for an $MA(1)$. We might naturally then ask whether there exists a zero-mean $MA(1)$ representation for Y,

$$Y_t = \varepsilon_t + \theta\varepsilon_{t-1}, \tag{4.7.7}$$

with

$$E(\varepsilon_t \varepsilon_{t-j}) = \begin{cases} \sigma^2 & \text{for } j = 0 \\ 0 & \text{otherwise,} \end{cases}$$

whose autocovariances match those implied by [4.7.6]. The autocovariances of [4.7.7] would be given by

$$E(Y_t Y_{t-j}) = \begin{cases} (1 + \theta^2)\sigma^2 & \text{for } j = 0 \\ \theta\sigma^2 & \text{for } j = \pm 1 \\ 0 & \text{otherwise.} \end{cases}$$

In order to be consistent with [4.7.6], it would have to be the case that

$$(1 + \theta^2)\sigma^2 = (1 + \delta^2)\sigma_u^2 + \sigma_v^2 \tag{4.7.8}$$

and

$$\theta\sigma^2 = \delta\sigma_u^2. \tag{4.7.9}$$

Equation [4.7.9] can be solved for σ^2,

$$\sigma^2 = \delta\sigma_u^2/\theta, \tag{4.7.10}$$

and then substituted into [4.7.8] to deduce

$$\begin{aligned} (1 + \theta^2)(\delta\sigma_u^2/\theta) &= (1 + \delta^2)\sigma_u^2 + \sigma_v^2 \\ (1 + \theta^2)\delta &= [(1 + \delta^2) + (\sigma_v^2/\sigma_u^2)]\theta \\ \delta\theta^2 - [(1 + \delta^2) + (\sigma_v^2/\sigma_u^2)]\theta + \delta &= 0. \end{aligned} \tag{4.7.11}$$

For given values of δ, σ_u^2, and σ_v^2, two values of θ that satisfy [4.7.11] can be found from the quadratic formula:

$$\theta = \frac{[(1 + \delta^2) + (\sigma_v^2/\sigma_u^2)] \pm \sqrt{[(1 + \delta^2) + (\sigma_v^2/\sigma_u^2)]^2 - 4\delta^2}}{2\delta}. \qquad [4.7.12]$$

If σ_v^2 were equal to zero, the quadratic equation in [4.7.11] would just be

$$\delta\theta^2 - (1 + \delta^2)\theta + \delta = \delta(\theta - \delta)(\theta - \delta^{-1}) = 0, \qquad [4.7.13]$$

whose solutions are $\theta = \delta$ and $\tilde{\theta} = \delta^{-1}$, the moving average parameter for X_t from the invertible and noninvertible representations, respectively. Figure 4.1 graphs equations [4.7.11] and [4.7.13] as functions of θ assuming positive autocorrelation for X_t ($\delta > 0$). For $\theta > 0$ and $\sigma_v^2 > 0$, equation [4.7.11] is everywhere lower than [4.7.13] by the amount $(\sigma_v^2/\sigma_u^2)\theta$, implying that [4.7.11] has two real solutions for θ, an invertible solution θ^* satisfying

$$0 < |\theta^*| < |\delta|, \qquad [4.7.14]$$

and a noninvertible solution $\tilde{\theta}^*$ characterized by

$$1 < |\delta^{-1}| < |\tilde{\theta}^*|.$$

Taking the values associated with the invertible representation (θ^*, σ^{*2}), let us consider whether [4.7.7] could indeed characterize the data $\{Y_t\}$ generated by [4.7.5]. This would require

$$(1 + \theta^* L)\varepsilon_t = (1 + \delta L)u_t + v_t, \qquad [4.7.15]$$

or

$$\begin{aligned} \varepsilon_t &= (1 + \theta^* L)^{-1}[(1 + \delta L)u_t + v_t] \\ &= (u_t - \theta^* u_{t-1} + \theta^{*2} u_{t-2} - \theta^{*3} u_{t-3} + \cdots) \\ &\quad + \delta(u_{t-1} - \theta^* u_{t-2} + \theta^{*2} u_{t-3} - \theta^{*3} u_{t-4} + \cdots) \\ &\quad + (v_t - \theta^* v_{t-1} + \theta^{*2} v_{t-2} - \theta^{*3} v_{t-3} + \cdots). \end{aligned} \qquad [4.7.16]$$

The series ε_t defined in [4.7.16] is a distributed lag on past values of u and v, so it might seem to possess a rich autocorrelation structure. In fact, it turns out to be

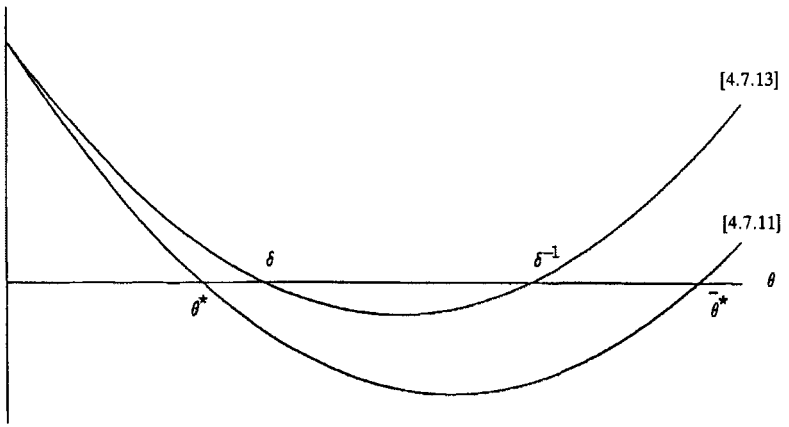


FIGURE 4.1 Graphs of equations [4.7.13] and [4.7.11].

white noise! To see this, note from [4.7.6] that the autocovariance-generating function of Y can be written

$$g_Y(z) = (1 + \delta z)\sigma_u^2(1 + \delta z^{-1}) + \sigma_v^2, \qquad [4.7.17]$$

so that the autocovariance-generating function of $\varepsilon_t = (1 + \theta^* L)^{-1} Y_t$ is

$$g_\varepsilon(z) = \frac{(1 + \delta z)\sigma_u^2(1 + \delta z^{-1}) + \sigma_v^2}{(1 + \theta^* z)(1 + \theta^* z^{-1})}. \qquad [4.7.18]$$

But θ^* and σ^{*2} were chosen so as to make the autocovariance-generating function of $(1 + \theta^* L)\varepsilon_t$, namely,

$$(1 + \theta^* z)\sigma^{*2}(1 + \theta^* z^{-1}),$$

identical to the right side of [4.7.17]. Thus, [4.7.18] is simply equal to

$$g_\varepsilon(z) = \sigma^{*2},$$

a white noise series.

To summarize, adding an $MA(1)$ process to a white noise series with which it is uncorrelated at all leads and lags produces a new $MA(1)$ process characterized by [4.7.7].

Note that the series ε_t in [4.7.16] could not be forecast as a linear function of lagged ε or of lagged Y. Clearly, ε could be forecast, however, on the basis of lagged u or lagged v. The histories $\{u_t\}$ and $\{v_t\}$ contain more information than $\{\varepsilon_t\}$ or $\{Y_t\}$. The optimal forecast of Y_{t+1} on the basis of $\{Y_t, Y_{t-1}, \ldots\}$ would be

$$\hat{E}(Y_{t+1}|Y_t, Y_{t-1}, \ldots) = \theta^* \varepsilon_t$$

with associated mean squared error σ^{*2}. By contrast, the optimal linear forecast of Y_{t+1} on the basis of $\{u_t, u_{t-1}, \ldots, v_t, v_{t-1}, \ldots\}$ would be

$$\hat{E}(Y_{t+1}|u_t, u_{t-1}, \ldots, v_t, v_{t-1}, \ldots) = \delta u_t$$

with associated mean squared error $\sigma_u^2 + \sigma_v^2$. Recalling from [4.7.14] that $|\theta^*| < |\delta|$, it appears from [4.7.9] that $(\theta^{*2})\sigma^{*2} < \delta^2\sigma_u^2$, meaning from [4.7.8] that $\sigma^2 > \sigma_u^2 + \sigma_v^2$. In other words, past values of Y contain less information than past values of u and v.

This example can be useful for thinking about the consequences of differing information sets. One can always make a sensible forecast on the basis of what one knows, $\{Y_t, Y_{t-1}, \ldots\}$, though usually there is other information that could have helped more. An important feature of such settings is that even though ε_t, u_t, and v_t are all white noise, there are complicated correlations between these white noise series.

Another point worth noting is that all that can be estimated on the basis of $\{Y_t\}$ are the two parameters θ^* and σ^{*2}, whereas the true "structural" model [4.7.5] has three parameters (δ, σ_u^2, and σ_v^2). Thus the parameters of the structural model are *unidentified* in the sense in which econometricians use this term—there exists a family of alternative configurations of δ, σ_u^2, and σ_v^2 with $|\delta| < 1$ that would produce the identical value for the likelihood function of the observed data $\{Y_t\}$.

The processes that were added together for this example both had mean zero. Adding constant terms to the processes will not change the results in any interesting way—if X_t is an $MA(1)$ process with mean μ_X and if v_t is white noise plus a constant μ_v, then $X_t + v_t$ will be an $MA(1)$ process with mean given by $\mu_X + \mu_v$. Thus, nothing is lost by restricting the subsequent discussion to sums of zero-mean processes.

Adding Two Moving Average Processes

Suppose next that X_t is a zero-mean $MA(q_1)$ process:

$$X_t = (1 + \delta_1 L + \delta_2 L^2 + \cdots + \delta_{q_1} L^{q_1})u_t \equiv \delta(L)u_t,$$

with

$$E(u_t u_{t-j}) = \begin{cases} \sigma_u^2 & \text{for } j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Let W_t be a zero-mean $MA(q_2)$ process:

$$W_t = (1 + \kappa_1 L + \kappa_2 L^2 + \cdots + \kappa_{q_2} L^{q_2})v_t \equiv \kappa(L)v_t,$$

with

$$E(v_t v_{t-j}) = \begin{cases} \sigma_v^2 & \text{for } j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, X has autocovariances $\gamma_0^X, \gamma_1^X, \ldots, \gamma_{q_1}^X$ of the form of [3.3.12] while W has autocovariances $\gamma_0^W, \gamma_1^W, \ldots, \gamma_{q_2}^W$ of the same basic structure. Assume that X and W are uncorrelated with each other at all leads and lags:

$$E(X_t W_{t-j}) = 0 \qquad \text{for all } j;$$

and suppose we observe

$$Y_t = X_t + W_t.$$

Define q to be the larger of q_1 or q_2:

$$q \equiv \max\{q_1, q_2\}.$$

Then the jth autocovariance of Y is given by

$$\begin{aligned} E(Y_t Y_{t-j}) &= E(X_t + W_t)(X_{t-j} + W_{t-j}) \\ &= E(X_t X_{t-j}) + E(W_t W_{t-j}) \\ &= \begin{cases} \gamma_j^X + \gamma_j^W & \text{for } j = 0, \pm 1, \pm 2, \ldots, \pm q \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Thus the autocovariances are zero beyond q lags, suggesting that Y_t might be represented as an $MA(q)$ process.

What more would we need to show to be fully convinced that Y_t is indeed an $MA(q)$ process? This question can be posed in terms of autocovariance-generating functions. Since

$$\gamma_j^Y = \gamma_j^X + \gamma_j^W,$$

it follows that

$$\sum_{j=-\infty}^{\infty} \gamma_j^Y z^j = \sum_{j=-\infty}^{\infty} \gamma_j^X z^j + \sum_{j=-\infty}^{\infty} \gamma_j^W z^j.$$

But these are just the definitions of the respective autocovariance-generating functions,

$$g_Y(z) = g_X(z) + g_W(z). \qquad [4.7.19]$$

Equation [4.7.19] is a quite general result—if one adds together two covariance-stationary processes that are uncorrelated with each other at all leads and lags, the

autocovariance-generating function of the sum is the sum of the autocovariance-generating functions of the individual series.

If Y_t is to be expressed as an $MA(q)$ process,

$$Y_t = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t \equiv \theta(L)\varepsilon_t$$

with

$$E(\varepsilon_t \varepsilon_{t-j}) = \begin{cases} \sigma^2 & \text{for } j = 0 \\ 0 & \text{otherwise,} \end{cases}$$

then its autocovariance-generating function would be

$$g_Y(z) = \theta(z)\theta(z^{-1})\sigma^2.$$

The question is thus whether there always exist values of $(\theta_1, \theta_2, \ldots, \theta_q, \sigma^2)$ such that [4.7.19] is satisfied:

$$\theta(z)\theta(z^{-1})\sigma^2 = \delta(z)\delta(z^{-1})\sigma_u^2 + \kappa(z)\kappa(z^{-1})\sigma_v^2. \qquad [4.7.20]$$

It turns out that there do. Thus, the conjecture turns out to be correct that if two moving average processes that are uncorrelated with each other at all leads and lags are added together, the result is a new moving average process whose order is the larger of the order of the original two series:

$$MA(q_1) + MA(q_2) = MA(\max\{q_1, q_2\}). \qquad [4.7.21]$$

A proof of this assertion, along with a constructive algorithm for achieving the factorization in [4.7.20], will be provided in Chapter 13.

Adding Two Autoregressive Processes

Suppose now that X_t and W_t are two $AR(1)$ processes:

$$(1 - \pi L)X_t = u_t \qquad [4.7.22]$$

$$(1 - \rho L)W_t = v_t, \qquad [4.7.23]$$

where u_t and v_t are each white noise with u_t uncorrelated with v_τ for all t and τ. Again suppose that we observe

$$Y_t = X_t + W_t$$

and want to forecast Y_{t+1} on the basis of its own lagged values.

If, by chance, X and W share the same autoregressive parameter, or

$$\pi = \rho,$$

then [4.7.22] could simply be added directly to [4.7.23] to deduce

$$(1 - \pi L)X_t + (1 - \pi L)W_t = u_t + v_t$$

or

$$(1 - \pi L)(X_t + W_t) = u_t + v_t.$$

But the sum $u_t + v_t$ is white noise (as a special case of result [4.7.21]), meaning that Y_t has an $AR(1)$ representation

$$(1 - \pi L)Y_t = \varepsilon_t.$$

In the more likely case that the autoregressive parameters π and ρ are different, then [4.7.22] can be multiplied by $(1 - \rho L)$:

$$(1 - \rho L)(1 - \pi L)X_t = (1 - \rho L)u_t; \qquad [4.7.24]$$

and similarly, [4.7.23] could be multiplied by $(1 - \pi L)$:

$$(1 - \pi L)(1 - \rho L)W_t = (1 - \pi L)v_t. \qquad [4.7.25]$$

Adding [4.7.24] to [4.7.25] produces

$$(1 - \rho L)(1 - \pi L)(X_t + W_t) = (1 - \rho L)u_t + (1 - \pi L)v_t. \qquad [4.7.26]$$

From [4.7.21], the right side of [4.7.26] has an $MA(1)$ representation. Thus, we could write

$$(1 - \phi_1 L - \phi_2 L^2)Y_t = (1 + \theta L)\varepsilon_t,$$

where

$$(1 - \phi_1 L - \phi_2 L^2) = (1 - \rho L)(1 - \pi L)$$

and

$$(1 + \theta L)\varepsilon_t = (1 - \rho L)u_t + (1 - \pi L)v_t.$$

In other words,

$$AR(1) + AR(1) = ARMA(2, 1). \qquad [4.7.27]$$

In general, adding an $AR(p_1)$ process

$$\pi(L)X_t = u_t,$$

to an $AR(p_2)$ process with which it is uncorrelated at all leads and lags,

$$\rho(L)W_t = v_t,$$

produces an $ARMA(p_1 + p_2, \max\{p_1, p_2\})$ process,

$$\phi(L)Y_t = \theta(L)\varepsilon_t,$$

where

$$\phi(L) = \pi(L)\rho(L)$$

and

$$\theta(L)\varepsilon_t = \rho(L)u_t + \pi(L)v_t.$$

4.8. *Wold's Decomposition and the Box-Jenkins Modeling Philosophy*

Wold's Decomposition

All of the covariance-stationary processes considered in Chapter 3 can be written in the form

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \qquad [4.8.1]$$

where ε_t is the white noise error one would make in forecasting Y_t as a linear function of lagged Y and where $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ with $\psi_0 = 1$.

One might think that we were able to write all these processes in the form of [4.8.1] because the discussion was restricted to a convenient class of models. However, the following result establishes that the representation [4.8.1] is in fact fundamental for any covariance-stationary time series.

Proposition 4.1: (Wold's decomposition). Any zero-mean covariance-stationary process Y_t can be represented in the form

$$Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \kappa_t, \qquad [4.8.2]$$

where $\psi_0 = 1$ and $\sum_{j=0}^{\infty} \psi_j^2 < \infty$. The term ε_t is white noise and represents the error made in forecasting Y_t on the basis of a linear function of lagged Y:

$$\varepsilon_t \equiv Y_t - \hat{E}(Y_t | Y_{t-1}, Y_{t-2}, \ldots). \qquad [4.8.3]$$

The value of κ_t is uncorrelated with ε_{t-j} for any j, though κ_t can be predicted arbitrarily well from a linear function of past values of Y:

$$\kappa_t = \hat{E}(\kappa_t | Y_{t-1}, Y_{t-2}, \ldots).$$

The term κ_t is called the *linearly deterministic* component of Y_t, while $\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$ is called the *linearly indeterministic* component. If $\kappa_t \equiv 0$, then the process is called *purely linearly indeterministic*.

This proposition was first proved by Wold (1938).[6] The proposition relies on stable second moments of Y but makes no use of higher moments. It thus describes only optimal linear forecasts of Y.

Finding the Wold representation in principle requires fitting an infinite number of parameters $(\psi_1, \psi_2, \ldots)$ to the data. With a finite number of observations on $(Y_1, Y_2, \ldots, Y_T)$, this will never be possible. As a practical matter, we therefore need to make some additional assumptions about the nature of $(\psi_1, \psi_2, \ldots)$. A typical assumption in Chapter 3 was that $\psi(L)$ can be expressed as the ratio of two finite-order polynomials:

$$\sum_{j=0}^{\infty} \psi_j L^j = \frac{\theta(L)}{\phi(L)} \equiv \frac{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q}{1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p}. \qquad [4.8.4]$$

Another approach, based on the presumed "smoothness" of the population spectrum, will be explored in Chapter 6.

The Box-Jenkins Modeling Philosophy

Many forecasters are persuaded of the benefits of parsimony, or using as few parameters as possible. Box and Jenkins (1976) have been influential advocates of this view. They noted that in practice, analysts end up replacing the true operators $\theta(L)$ and $\phi(L)$ with estimates $\hat{\theta}(L)$ and $\hat{\phi}(L)$ based on the data. The more parameters to estimate, the more room there is to go wrong.

Although complicated models can track the data very well over the historical period for which parameters are estimated, they often perform poorly when used for out-of-sample forecasting. For example, the 1960s saw the development of a number of large macroeconometric models purporting to describe the economy using hundreds of macroeconomic variables and equations. Part of the disillusionment with such efforts was the discovery that univariate *ARMA* models with small values of p or q often produced better forecasts than the big models (see for example Nelson, 1972).[7] As we shall see in later chapters, large size alone was hardly the only liability of these large-scale macroeconometric models. Even so, the claim that simpler models provide more robust forecasts has a great many believers across disciplines.

[6]See Sargent (1987, pp. 286–90) for a nice sketch of the intuition behind this result.

[7]For more recent pessimistic evidence about current large-scale models, see Ashley (1988).

The approach to forecasting advocated by Box and Jenkins can be broken down into four steps:

(1) Transform the data, if necessary, so that the assumption of covariance-stationarity is a reasonable one.
(2) Make an initial guess of small values for p and q for an $ARMA(p, q)$ model that might describe the transformed series.
(3) Estimate the parameters in $\phi(L)$ and $\theta(L)$.
(4) Perform diagnostic analysis to confirm that the model is indeed consistent with the observed features of the data.

The first step, selecting a suitable transformation of the data, is discussed in Chapter 15. For now we merely remark that for economic series that grow over time, many researchers use the change in the natural logarithm of the raw data. For example, if X_t is the level of real GNP in year t, then

$$Y_t = \log X_t - \log X_{t-1} \qquad [4.8.5]$$

might be the variable that an $ARMA$ model purports to describe.

The third and fourth steps, estimation and diagnostic testing, will be discussed in Chapters 5 and 14. Analysis of seasonal dynamics can also be an important part of step 2 of the procedure; this is briefly discussed in Section 6.4. The remainder of this section is devoted to an exposition of the second step in the Box-Jenkins procedure on nonseasonal data, namely, selecting candidate values for p and q.[8]

Sample Autocorrelations

An important part of this selection procedure is to form an estimate $\hat{\rho}_j$ of the population autocorrelation ρ_j. Recall that ρ_j was defined as

$$\rho_j \equiv \gamma_j/\gamma_0$$

where

$$\gamma_j = E(Y_t - \mu)(Y_{t-j} - \mu).$$

A natural estimate of the population autocorrelation ρ_j is provided by the corresponding sample moments:

$$\hat{\rho}_j = \hat{\gamma}_j/\hat{\gamma}_0,$$

where

$$\hat{\gamma}_j = \frac{1}{T}\sum_{t=j+1}^{T}(y_t - \bar{y})(y_{t-j} - \bar{y}) \qquad \text{for } j = 0, 1, 2, \ldots, T-1 \qquad [4.8.6]$$

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t. \qquad [4.8.7]$$

Note that even though only $T - j$ observations are used to construct $\hat{\gamma}_j$, the denominator in [4.8.6] is T rather than $T - j$. Thus, for large j, expression [4.8.6] shrinks the estimates toward zero, as indeed the population autocovariances go to zero as $j \to \infty$, assuming covariance-stationarity. Also, the full sample of observations is used to construct $\bar{y}$.

[8]Box and Jenkins refer to this step as "identification" of the appropriate model. We avoid Box and Jenkins's terminology, because "identification" has a quite different meaning for econometricians.

Recall that if the data really follow an $MA(q)$ process, then ρ_j will be zero for $j > q$. By contrast, if the data follow an $AR(p)$ process, then ρ_j will gradually decay toward zero as a mixture of exponentials or damped sinusoids. One guide for distinguishing between MA and AR representations, then, would be the decay properties of ρ_j. Often, we are interested in a quick assessment of whether $\rho_j = 0$ for $j = q + 1, q + 2, \ldots$. If the data were really generated by a Gaussian $MA(q)$ process, then the variance of the estimate $\hat{\rho}_j$ could be approximated by[9]

$$\text{Var}(\hat{\rho}_j) \simeq \frac{1}{T}\left\{1 + 2\sum_{i=1}^{q}\rho_i^2\right\} \qquad \text{for } j = q + 1, q + 2, \ldots. \qquad [4.8.8]$$

Thus, in particular, if we suspect that the data were generated by Gaussian white noise, then $\hat{\rho}_j$ for any $j \neq 0$ should lie between $\pm 2/\sqrt{T}$ about 95% of the time.

In general, if there is autocorrelation in the process that generated the original data $\{Y_t\}$, then the estimate $\hat{\rho}_j$ will be correlated with $\hat{\rho}_i$ for $i \neq j$.[10] Thus patterns in the estimated $\hat{\rho}_j$ may represent sampling error rather than patterns in the true ρ_j.

Partial Autocorrelation

Another useful measure is the *partial autocorrelation*. The mth population partial autocorrelation (denoted $\alpha_m^{(m)}$) is defined as the last coefficient in a linear projection of Y on its m most recent values (equation [4.3.7]):

$$\hat{Y}_{t+1|t} - \mu = \alpha_1^{(m)}(Y_t - \mu) + \alpha_2^{(m)}(Y_{t-1} - \mu) + \cdots + \alpha_m^{(m)}(Y_{t-m+1} - \mu).$$

We saw in equation [4.3.8] that the vector $\boldsymbol{\alpha}^{(m)}$ can be calculated from

$$\begin{bmatrix} \alpha_1^{(m)} \\ \alpha_2^{(m)} \\ \vdots \\ \alpha_m^{(m)} \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{m-2} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{bmatrix}.$$

Recall that if the data were really generated by an $AR(p)$ process, only the p most recent values of Y would be useful for forecasting. In this case, the projection coefficients on Y's more than p periods in the past are equal to zero:

$$\alpha_m^{(m)} = 0 \qquad \text{for } m = p + 1, p + 2, \ldots.$$

By contrast, if the data really were generated by an $MA(q)$ process with $q \geq 1$, then the partial autocorrelation $\alpha_m^{(m)}$ asymptotically approaches zero instead of cutting off abruptly.

A natural estimate of the mth partial autocorrelation is the last coefficient in an *OLS* regression of y on a constant and its m most recent values:

$$y_{t+1} = \hat{c} + \hat{\alpha}_1^{(m)}y_t + \hat{\alpha}_2^{(m)}y_{t-1} + \cdots + \hat{\alpha}_m^{(m)}y_{t-m+1} + \hat{e}_t,$$

where $\hat{e}_t$ denotes the *OLS* regression residual. If the data were really generated by an $AR(p)$ process, then the sample estimate $(\hat{\alpha}_m^{(m)})$ would have a variance around the true value (0) that could be approximated by[11]

$$\text{Var}(\hat{\alpha}_m^{(m)}) \simeq 1/T \qquad \text{for } m = p + 1, p + 2, \ldots.$$

[9]See Box and Jenkins (1976, p. 35).

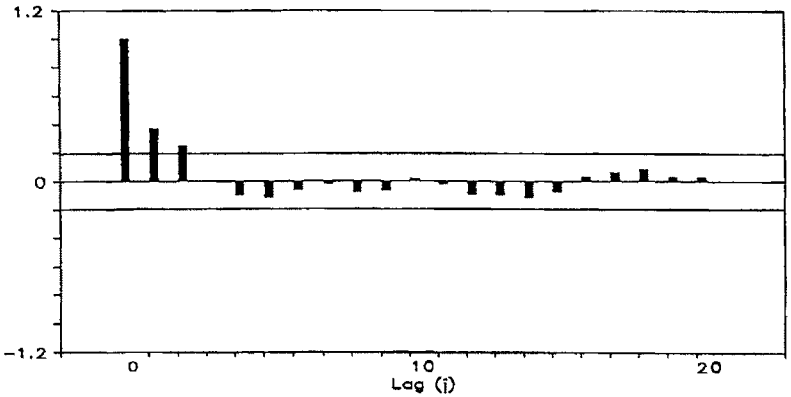
[10]Again, see Box and Jenkins (1976, p. 35).

[11]Box and Jenkins (1976, p. 65).

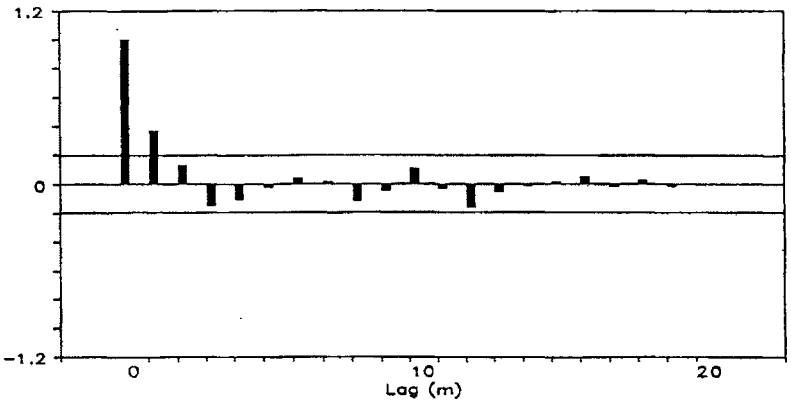
Moreover, if the data were really generated by an $AR(p)$ process, then $\hat{\alpha}_i^{(i)}$ and $\hat{\alpha}_j^{(j)}$ would be asymptotically independent for $i, j > p$.

Example 4.1

We illustrate the Box-Jenkins approach with seasonally adjusted quarterly data on U.S. real GNP from 1947 through 1988. The raw data (x_t) were converted to log changes (y_t) as in [4.8.5]. Panel (a) of Figure 4.2 plots the sample autocorrelations of y ($\hat{\rho}_j$ for $j = 0, 1, \ldots, 20$), while panel (b) displays the sample partial autocorrelations ($\hat{\alpha}_m^{(m)}$ for $m = 0, 1, \ldots, 20$). Ninety-five percent confidence bands ($\pm 2/\sqrt{T}$) are plotted on both panels; for panel (a), these are appropriate under the null hypothesis that the data are really white noise, whereas for panel (b) these are appropriate if the data are really generated by an $AR(p)$ process for p less than m.



(a) Sample autocorrelations



(b) Sample partial autocorrelations

FIGURE 4.2 Sample autocorrelations and partial autocorrelations for U.S. quarterly real GNP growth, 1947:II to 1988:IV. Ninety-five percent confidence intervals are plotted as $\pm 2/\sqrt{T}$.

The first two autocorrelations appear nonzero, suggesting that $q = 2$ would be needed to describe these data as coming from a moving average process. On the other hand, the pattern of autocorrelations appears consistent with the simple geometric decay of an $AR(1)$ process,

$$\rho_j = \phi^j$$

with $\phi \cong 0.4$. The partial autocorrelation could also be viewed as dying out after one lag, also consistent with the $AR(1)$ hypothesis. Thus, one's initial guess for a parsimonious model might be that GNP growth follows an $AR(1)$ process, with $MA(2)$ as another possibility to be considered.

APPENDIX 4.A. *Parallel Between* OLS *Regression and Linear Projection*

This appendix discusses the parallel between ordinary least squares regression and linear projection. This parallel is developed by introducing an artificial random variable specifically constructed so as to have population moments identical to the sample moments of a particular sample. Say that in some particular sample on which we intend to perform *OLS* we have observed T particular values for the explanatory vector, denoted $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$. Consider an artificial discrete-valued random variable ξ that can take on only one of these particular T values, each with probability $(1/T)$:

$$\begin{aligned} P\{\xi = \mathbf{x}_1\} &= 1/T \\ P\{\xi = \mathbf{x}_2\} &= 1/T \\ &\vdots \\ P\{\xi = \mathbf{x}_T\} &= 1/T. \end{aligned}$$

Thus ξ is an artificially constructed random variable whose population probability distribution is given by the empirical distribution function of $\mathbf{x}_t$. The population mean of the random variable ξ is

$$E(\xi) = \sum_{t=1}^{T} \mathbf{x}_t \cdot P\{\xi = \mathbf{x}_t\} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t.$$

Thus, the population mean of ξ equals the observed sample mean of the true random variable $\mathbf{X}_t$. The population second moment of ξ is

$$E(\xi\xi') = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t', \qquad [4.A.1]$$

which is the sample second moment of $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$.

We can similarly construct a second artificial variable ω that can take on one of the discrete values $(y_2, y_3, \ldots, y_{T+1})$. Suppose that the joint distribution of ω and ξ is given by

$$P\{\xi = \mathbf{x}_t, \omega = y_{t+1}\} = 1/T \qquad \text{for } t = 1, 2, \ldots, T.$$

Then

$$E(\xi\omega) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t y_{t+1}. \qquad [4.A.2]$$

The coefficient for a linear projection of ω on ξ is the value of $\boldsymbol{\alpha}$ that minimizes

$$E(\omega - \boldsymbol{\alpha}'\xi)^2 = \frac{1}{T}\sum_{t=1}^{T} (y_{t+1} - \boldsymbol{\alpha}'\mathbf{x}_t)^2. \qquad [4.A.3]$$

This is algebraically the same problem as choosing $\boldsymbol{\beta}$ so as to minimize [4.1.17]. Thus, ordinary least squares regression (choosing $\boldsymbol{\beta}$ so as to minimize [4.1.17]) can be viewed as a special case of linear projection (choosing $\boldsymbol{\alpha}$ so as to minimize [4.A.3]). The value of $\boldsymbol{\alpha}$

that minimizes [4.A.3] can be found from substituting the expressions for the population moments of the artificial random variables (equations [4.A.1] and [4.A.2]) into the formula for a linear projection (equation [4.1.13]):

$$\boldsymbol{\alpha} = [E(\boldsymbol{\xi}\boldsymbol{\xi}')]^{-1}E(\boldsymbol{\xi}\omega) = \left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t y_{t+1}\right].$$

Thus the formula for the *OLS* estimate $\mathbf{b}$ in [4.1.18] can be obtained as a special case of the formula for the linear projection coefficient $\boldsymbol{\alpha}$ in [4.1.13].

Because linear projections and *OLS* regressions share the same mathematical structure, statements about one have a parallel in the other. This can be a useful device for remembering results or confirming algebra. For example, the statement about population moments,

$$E(Y^2) = \text{Var}(Y) + [E(Y)]^2, \qquad [4.A.4]$$

has the sample analog

$$\frac{1}{T}\sum_{t=1}^{T} y_t^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2 + (\bar{y})^2 \qquad [4.A.5]$$

with $\bar{y} = (1/T)\sum_{t=1}^{T} y_t$.

As a second example, suppose that we estimate a series of n *OLS* regressions, with y_{it} the dependent variable for the ith regression and $\mathbf{x}_t$ a $(k \times 1)$ vector of explanatory variables common to each regression. Let $\mathbf{y}_t \equiv (y_{1t}, y_{2t}, \ldots, y_{nt})'$ and write the regression model as

$$\mathbf{y}_t = \boldsymbol{\Pi}'\mathbf{x}_t + \mathbf{u}_t$$

for $\boldsymbol{\Pi}'$ an $(n \times k)$ matrix of regression coefficients. Then the sample variance-covariance matrix of the *OLS* residuals can be inferred from [4.1.24]:

$$\frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{u}}_t\hat{\mathbf{u}}_t' = \left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}_t'\right] - \left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{x}_t'\right]\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{y}_t'\right], \qquad [4.A.6]$$

where $\hat{\mathbf{u}}_t = \mathbf{y}_t - \hat{\boldsymbol{\Pi}}'\mathbf{x}_t$, and the ith row of $\hat{\boldsymbol{\Pi}}'$ is given by

$$\hat{\boldsymbol{\pi}}_i' = \left\{\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t y_{it}\right]\right\}'.$$

APPENDIX 4.B. *Triangular Factorization of the Covariance Matrix for an* MA(1) *Process*

This appendix establishes that the triangular factorization of $\boldsymbol{\Omega}$ in [4.5.17] is given by [4.5.18] and [4.5.19].

The magnitude σ^2 is simply a constant term that will end up multiplying every term in the $\mathbf{D}$ matrix. Recognizing this, we can initially solve the factorization assuming that $\sigma^2 = 1$, and then multiply the resulting $\mathbf{D}$ matrix by σ^2 to obtain the result for the general case. The (1, 1) element of $\mathbf{D}$ (ignoring the factor σ^2) is given by the (1, 1) element of $\boldsymbol{\Omega}$: $d_{11} = (1 + \theta^2)$. To put a zero in the (2, 1) position of $\boldsymbol{\Omega}$, we multiply the first row of $\boldsymbol{\Omega}$ by $\theta/(1 + \theta^2)$ and subtract the result from the second; hence, $a_{21} = \theta/(1 + \theta^2)$. This operation changes the (2, 2) element of $\boldsymbol{\Omega}$ to

$$d_{22} = (1 + \theta^2) - \frac{\theta^2}{1 + \theta^2} = \frac{(1 + \theta^2)^2 - \theta^2}{1 + \theta^2} = \frac{1 + \theta^2 + \theta^4}{1 + \theta^2}.$$

To put a zero in the (3, 2) element of $\boldsymbol{\Omega}$, the second row of the new matrix must be multiplied by θ/d_{22} and then subtracted from the third row; hence,

$$a_{32} = \theta/d_{22} = \frac{\theta(1 + \theta^2)}{1 + \theta^2 + \theta^4}.$$

This changes the (3, 3) element to

$$\begin{aligned} d_{33} &= (1 + \theta^2) - \frac{\theta^2(1 + \theta^2)}{1 + \theta^2 + \theta^4} \\ &= \frac{(1 + \theta^2)(1 + \theta^2 + \theta^4) - \theta^2(1 + \theta^2)}{1 + \theta^2 + \theta^4} \\ &= \frac{(1 + \theta^2 + \theta^4) + \theta^2(1 + \theta^2 + \theta^4) - \theta^2(1 + \theta^2)}{1 + \theta^2 + \theta^4} \\ &= \frac{1 + \theta^2 + \theta^4 + \theta^6}{1 + \theta^2 + \theta^4}. \end{aligned}$$

In general, for the ith row,

$$d_{ii} = \frac{1 + \theta^2 + \theta^4 + \cdots + \theta^{2i}}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2(i-1)}}.$$

To put a zero in the $(i + 1, i)$ position, multiply by

$$a_{i+1,i} = \theta/d_{ii} = \frac{\theta[1 + \theta^2 + \theta^4 + \cdots + \theta^{2(i-1)}]}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2i}}$$

and subtract from the $(i + 1)$th row, producing

$$\begin{aligned} d_{i+1,i+1} &= (1 + \theta^2) - \frac{\theta^2[1 + \theta^2 + \theta^4 + \cdots + \theta^{2(i-1)}]}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2i}} \\ &= \frac{(1 + \theta^2 + \theta^4 + \cdots + \theta^{2i}) + \theta^2(1 + \theta^2 + \theta^4 + \cdots + \theta^{2i})}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2i}} \\ &\quad - \frac{\theta^2[1 + \theta^2 + \theta^4 + \cdots + \theta^{2(i-1)}]}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2i}} \\ &= \frac{1 + \theta^2 + \theta^4 + \cdots + \theta^{2(i+1)}}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2i}}. \end{aligned}$$

Chapter 4 Exercises

4.1. Use formula [4.3.6] to show that for a covariance-stationary process, the projection of Y_{t+1} on a constant and Y_t is given by

$$\hat{E}(Y_{t+1}|Y_t) = (1 - \rho_1)\mu + \rho_1 Y_t$$

where $\mu = E(Y_t)$ and $\rho_1 = \gamma_1/\gamma_0$.

(a) Show that for the $AR(1)$ process, this reproduces equation [4.2.19] for $s = 1$.
(b) Show that for the $MA(1)$ process, this reproduces equation [4.5.20] for $n = 2$.
(c) Show that for an $AR(2)$ process, the implied forecast is

$$\mu + [\phi_1/(1 - \phi_2)](Y_t - \mu).$$

Is the error associated with this forecast correlated with Y_t? Is it correlated with Y_{t-1}?

4.2. Verify equation [4.3.3].

4.3. Find the triangular factorization of the following matrix:

$$\begin{bmatrix} 1 & -2 & 3 \\ -2 & 6 & -4 \\ 3 & -4 & 12 \end{bmatrix}.$$

4.4. Can the coefficient on Y_2 from a linear projection of Y_4 on Y_3, Y_2, and Y_1 be found from the (4, 2) element of the matrix $\mathbf{A}$ from the triangular factorization of $\mathbf{\Omega} = E(\mathbf{YY}')$?

4.5. Suppose that X_t follows an $AR(p)$ process and v_t is a white noise process that is uncorrelated with X_{t-j} for all j. Show that the sum

$$Y_t = X_t + v_t$$

follows an $ARMA(p, p)$ process.

4.6. Generalize Exercise 4.5 to deduce that if one adds together an $AR(p)$ process with an $MA(q)$ process and if these two processes are uncorrelated with each other at all leads and lags, then the result is an $ARMA(p, p + q)$ process.

Chapter 4 References

Ashley, Richard. 1988. "On the Relative Worth of Recent Macroeconomic Forecasts." *International Journal of Forecasting* 4:363–76.

Box, George E. P., and Gwilym M. Jenkins. 1976. *Time Series Analysis: Forecasting and Control*, rev. ed. San Francisco: Holden-Day.

Nelson, Charles R. 1972. "The Prediction Performance of the F.R.B.–M.I.T.–PENN Model of the U.S. Economy." *American Economic Review* 62:902–17.

Sargent, Thomas J. 1987. *Macroeconomic Theory*, 2d ed. Boston: Academic Press.

Wold, Herman. 1938 (2d ed. 1954). *A Study in the Analysis of Stationary Time Series*. Uppsala, Sweden: Almqvist and Wiksell.

5 Maximum Likelihood Estimation

5.1. *Introduction*

Consider an *ARMA* model of the form

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \quad [5.1.1]$$

with ε_t white noise:

$$E(\varepsilon_t) = 0 \quad [5.1.2]$$

$$E(\varepsilon_t \varepsilon_\tau) = \begin{cases} \sigma^2 & \text{for } t = \tau \\ 0 & \text{otherwise.} \end{cases} \quad [5.1.3]$$

The previous chapters assumed that the population parameters $(c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)$ were known and showed how population moments such as $E(Y_t Y_{t-j})$ and linear forecasts $\hat{E}(Y_{t+s} | Y_t, Y_{t-1}, \ldots)$ could be calculated as functions of these population parameters. This chapter explores how to estimate the values of $(c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)$ on the basis of observations on Y.

The primary principle on which estimation will be based is *maximum likelihood*. Let $\boldsymbol{\theta} \equiv (c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)'$ denote the vector of population parameters. Suppose we have observed a sample of size T $(y_1, y_2, \ldots, y_T)$. The approach will be to calculate the probability density

$$f_{Y_T, Y_{T-1}, \ldots, Y_1}(y_T, y_{T-1}, \ldots, y_1; \boldsymbol{\theta}), \quad [5.1.4]$$

which might loosely be viewed as the probability of having observed this particular sample. The maximum likelihood estimate (*MLE*) of $\boldsymbol{\theta}$ is the value for which this sample is most likely to have been observed; that is, it is the value of $\boldsymbol{\theta}$ that maximizes [5.1.4].

This approach requires specifying a particular distribution for the white noise process ε_t. Typically we will assume that ε_t is Gaussian white noise:

$$\varepsilon_t \sim \text{i.i.d. } N(0, \sigma^2). \quad [5.1.5]$$

Although this assumption is strong, the estimates of $\boldsymbol{\theta}$ that result from it will often turn out to be sensible for non-Gaussian processes as well.

Finding maximum likelihood estimates conceptually involves two steps. First, the likelihood function [5.1.4] must be calculated. Second, values of $\boldsymbol{\theta}$ must be found that maximize this function. This chapter is organized around these two steps. Sections 5.2 through 5.6 show how to calculate the likelihood function for different Gaussian *ARMA* specifications, while subsequent sections review general techniques for numerical optimization.