# 13 : CHAPTER

# Linear Stochastic Models

### Stationary Stochastic processes

A temporal stochastic process is simply a sequence of random variables indexed by a time subscript. Such a process can be denoted by $x(t)$. The element of the sequence at the point $t = \tau$ is $x_\tau = x(\tau)$.

Let $x(\tau) = [x_{\tau+1}, x_{\tau+2}, \ldots, x_{\tau+n}]'$ denote $n$ consecutive elements of the sequence. Then the process is said to be strictly stationary if the joint probability distribution of the elements does not depend on $\tau$ regardless of the size of $n$. This means that any two segments of the sequence of equal length have identical probability density functions. In consequence, the decision on where to place the time origin is arbitrary; and the argument $\tau$ can be omitted. Some further implications of stationarity are that

$$E(x_t) = \mu < \infty \quad \text{for all} \quad t \quad \text{and} \quad C(x_{\tau+t}, x_{\tau+s}) = \gamma_{|t-s|}. \tag{1}$$

The latter condition means that the covariance of any two elements depends only on their temporal separation $|t - s|$. Notice that, if the elements of the sequence are normally distributed, then the two conditions are sufficient to establish strict stationarity. On their own, they constitute the conditions of weak or 2nd-order stationarity.

The condition on the covariances implies that the dispersion matrix of the vector $x = [x_1, x_2, \ldots, x_n]'$ is a bisymmetric Laurent matrix of the form

$$\begin{aligned} D(x) &= E\big\{[x - E(x)][x - E(x)]'\big\} \\ &= \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{n-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \gamma_{n-3} & \cdots & \gamma_0 \end{bmatrix}. \end{aligned} \tag{2}$$

Given that a sequence of observations of a time series represents only a segment of a single realisation of a stochastic process, one might imagine that

there is little chance of making valid inferences about the parameters of the process. However, provided that the process $x(t)$ is stationary and provided that the statistical dependencies between widely separated elements of the sequence are weak, it is possible to estimate consistently those parameters of the process which express the dependence of proximate elements of the sequence. If one is prepared to make sufficiently strong assumptions about the nature of the process, then a knowledge of such parameters may be all that is needed for a complete characterisation of the process.

**Moving Average Processes**

The $q$th-order moving average process, or MA($q$) process, is defined by the equation

$$y(t) = \mu_0\varepsilon(t) + \mu_1\varepsilon(t-1) + \cdots + \mu_q\varepsilon(t-q), \tag{3}$$

where $\varepsilon(t)$, which has $E\{x(t)\} = 0$, is a white-noise process consisting of a sequence of independently and identically distributed random variables with zero expectations. The equation is normalised either by setting $\mu_0 = 1$ or by setting $V\{\varepsilon(t)\} = \sigma_\varepsilon^2 = 1$. The equation can be written in summary notation as $y(t) = \mu(L)\varepsilon(t)$, where $\mu(L) = \mu_0 + \mu_1 L + \cdots + \mu_q L^q$ is a polynomial in the lag operator.

A moving-average process is clearly stationary since any two elements $y_t$ and $y_s$ represent the same function of identically distributed vectors $\varepsilon_t = [\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}]'$ and $\varepsilon_s = [\varepsilon_s, \varepsilon_{s-1}, \ldots, \varepsilon_{s-q}]'$ . In addition to the condition of stationarity, it is usually required that a moving-average process should be invertible such that it can be expressed in the form of $\mu^{-1}(L)y(t) = \varepsilon(t)$ where the LHS embodies a convergent sum of past values of $y(t)$. This is an infinite-order autoregressive representation of the process. The representation is available only if all the roots of the equation $\mu(z) = \mu_0 + \mu_1 z + \cdots + \mu_q z^q = 0$ lie outside the unit circle. This conclusion follows from our discussion of partial fractions.

As an example, let us consider the first-order moving-average process which is defined by

$$y(t) = \varepsilon(t) - \theta\varepsilon(t-1) = (1 - \theta L)\varepsilon(t). \tag{4}$$

Provided that $|\theta| < 1$, this can be written in autoregressive form as

$$\begin{aligned} \varepsilon(t) &= (1 - \theta L)^{-1} y(t) \\ &= \big\{y(t) + \theta y(t-1) + \theta^2 y(t-2) + \cdots\big\}. \end{aligned} \tag{5}$$

Imagine that $|\theta| > 1$ instead. Then, to obtain a convergent series, we have to write

$$\begin{aligned} y(t+1) &= \varepsilon(t+1) - \theta\varepsilon(t) \\ &= -\theta(1 - L^{-1}/\theta)\varepsilon(t), \end{aligned} \tag{6}$$

where $L^{-1}\varepsilon(t) = \varepsilon(t+1)$. This gives

$$\begin{aligned}\varepsilon(t) &= -\theta^{-1}(1 - L^{-1}/\theta)^{-1}y(t+1)\\ &= -\theta^{-1}\big\{y(t+1)/\theta + y(t+2)/\theta^2 + y(t-3)/\theta^3 + \cdots\big\}.\end{aligned} \tag{7}$$

Normally, an expression such as this, which embodies future values of $y(t)$, would have no reasonable meaning.

It is straightforward to generate the sequence of autocovariances from a knowledge of the parameters of the moving-average process and of the variance of the white-noise process. Consider

$$\begin{aligned}\gamma_\tau &= E(y_t y_{t-\tau})\\ &= E\Big\{\sum_i \mu_i\varepsilon_{t-i}\sum_j \mu_j\varepsilon_{t-\tau-j}\Big\}\\ &= \sum_i\sum_j \mu_i\mu_j E(\varepsilon_{t-i}\varepsilon_{t-\tau-j}).\end{aligned} \tag{8}$$

Since $\varepsilon(t)$ is a sequence of independently and identically distributed random variables with zero expectations, it follows that

$$E(\varepsilon_{t-i}\varepsilon_{t-\tau-j}) = \begin{cases}0, & \text{if } i \neq \tau + j;\\ \sigma_\varepsilon^2, & \text{if } i = \tau + j.\end{cases} \tag{9}$$

Therefore

$$\gamma_\tau = \sigma_\varepsilon^2 \sum_j \mu_j\mu_{j+\tau}. \tag{10}$$

Now let $\tau = 0, 1, \ldots, q$. This gives

$$\begin{aligned}\gamma_0 &= \sigma_\varepsilon^2(\mu_0^2 + \mu_1^2 + \cdots + \mu_q^2),\\ \gamma_1 &= \sigma_\varepsilon^2(\mu_0\mu_1 + \mu_1\mu_2 + \cdots + \mu_{q-1}\mu_q),\\ &\vdots\\ \gamma_q &= \sigma_\varepsilon^2\mu_0\mu_q.\end{aligned} \tag{11}$$

Also, $\gamma_\tau = 0$ for all $\tau > q$.

The first-order moving-average process $y(t) = \varepsilon(t) - \theta\varepsilon(t-1)$ has the following autocovariances:

$$\begin{aligned}\gamma_0 &= \sigma_\varepsilon^2(1 + \theta^2),\\ \gamma_1 &= -\sigma_\varepsilon^2\theta,\\ \gamma_\tau &= 0 \quad \text{if} \quad \tau > 1.\end{aligned} \tag{12}$$

Thus, for a vector $y = [y_1, y_2, \ldots, y_T]'$ of $T$ consecutive elements from a first-order moving-average process, the dispersion matrix is

$$D(y) = \sigma_\varepsilon^2 \begin{bmatrix} 1+\theta^2 & -\theta & 0 & \ldots & 0 \\ -\theta & 1+\theta^2 & -\theta & \ldots & 0 \\ 0 & -\theta & 1+\theta^2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1+\theta^2 \end{bmatrix}. \tag{13}$$

In general, the dispersion matrix of a $q$th-order moving-average process has $q$ subdiagonal and $q$ supradiagonal bands of nonzero elements and zero elements elsewhere.

It is also helpful to define an autocovariance generating function which is a power series whose coefficients are the autocovariances $\gamma_\tau$ for successive values of $\tau$. This is denoted by

$$\gamma(z) = \sum_\tau \gamma_\tau z^\tau; \quad \text{with} \quad \tau = \{0, \pm 1, \pm 2, \ldots\} \quad \text{and} \quad \gamma_\tau = \gamma_{-\tau}. \tag{14}$$

The generating function is also called the $z$-transform of the autocovariance function.

The autocovariance generating function of the $q$th-order moving-average process can be found quite readily. Consider the convolution

$$\begin{aligned} \mu(z)\mu(z^{-1}) &= \sum_i \mu_i z^i \sum_j \mu_j z^{-j} \\ &= \sum_i \sum_j \mu_i \mu_j z^{i-j} \\ &= \sum_\tau \Big( \sum_j \mu_i \mu_{j+\tau} \Big) z^\tau, \quad \tau = i - j. \end{aligned} \tag{15}$$

By referring to the expression for the autocovariance of lag $\tau$ of a moving-average process given under (10), it can be seen that the autocovariance generating function is just

$$\gamma(z) = \sigma_\varepsilon^2 \mu(z)\mu(z^{-1}). \tag{16}$$

**Autoregressive Processes**

The $p$th-order autoregressive process, or AR($p$) process, is defined by the equation

$$\alpha_0 y(t) + \alpha_1 y(t-1) + \cdots + \alpha_p y(t-p) = \varepsilon(t). \tag{17}$$

This equation is invariably normalised by setting $\alpha_0 = 1$, although it would be possible to set $\sigma_\varepsilon^2 = 1$ instead. The equation can be written in summary notation as $\alpha(L)y(t) = \varepsilon(t)$, where $\alpha(L) = \alpha_0 + \alpha_1 L + \cdots + \alpha_p L^p$. For the process to be stationary, the roots of the equation $\alpha(z) = \alpha_0 + \alpha_1 z + \cdots + \alpha_p z^p = 0$ must lie outside the unit circle. This condition enables us to write the autoregressive process as an infinite-order moving-average process in the form of $y(t) = \alpha^{-1}(L)\varepsilon(t)$.

As an example, let us consider the first-order autoregressive process which is defined by

$$\begin{aligned} \varepsilon(t) &= y(t) - \phi y(t-1) \\ &= (1 - \phi L)y(t). \end{aligned} \tag{18}$$

Provided that the process is stationary with $|\phi| < 1$, it can be represented in moving-average form as

$$\begin{aligned} y(t) &= (1 - \phi L)^{-1}\varepsilon(t) \\ &= \big\{\varepsilon(t) + \phi\varepsilon(t-1) + \phi^2\varepsilon(t-2) + \cdots\big\}. \end{aligned} \tag{19}$$

The autocovariances of the process can be found by using the formula of (10) which is applicable to moving-average process of finite or infinite order. Thus

$$\begin{aligned} \gamma_\tau &= E(y_t y_{t-\tau}) \\ &= E\Big\{\sum_i \phi^i \varepsilon_{t-i} \sum_j \phi^j \varepsilon_{t-\tau-j}\Big\} \\ &= \sum_i \sum_j \phi^i \phi^j E(\varepsilon_{t-i}\varepsilon_{t-\tau-j}); \end{aligned} \tag{20}$$

and the result under (9) indicates that

$$\begin{aligned} \gamma_\tau &= \sigma_\varepsilon^2 \sum_j \phi^j \phi^{j+\tau} \\ &= \frac{\sigma_\varepsilon^2 \phi^\tau}{1 - \phi^2}. \end{aligned} \tag{21}$$

For a vector $y = [y_1, y_2, \ldots, y_T]'$ of $T$ consecutive elements from a first-order autoregressive process, the dispersion matrix has the form

$$D(y) = \frac{\sigma_\varepsilon^2}{1 - \phi^2} \begin{bmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{T-1} \\ \phi & 1 & \phi & \cdots & \phi^{T-2} \\ \phi^2 & \phi & 1 & \cdots & \phi^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \cdots & 1 \end{bmatrix}. \tag{22}$$

To find the autocovariance generating function for the general $p$th-order autoregressive process, we may consider again the function $\alpha(z) = \sum_i \alpha_i z^i$. Since an autoregressive process may be treated as an infinite-order moving-average process, it follows that

$$\gamma(z) = \frac{\sigma_\varepsilon^2}{\alpha(z)\alpha(z^{-1})}. \tag{23}$$

For an alternative way of finding the autocovariances of the $p$th-order process, consider multiplying $\sum_i \alpha_i y_{t-i} = \varepsilon_t$ by $y_{t-\tau}$ and taking expectations to give

$$\sum_i \alpha_i E(y_{t-i}y_{t-\tau}) = E(\varepsilon_t y_{t-\tau}). \tag{24}$$

Taking account of the normalisation $\alpha_0 = 1$, we find that

$$E(\varepsilon_t y_{t-\tau}) = \begin{cases} \sigma_\varepsilon^2, & \text{if } \tau = 0; \\ 0, & \text{if } \tau > 0. \end{cases} \tag{25}$$

Therefore, on setting $E(y_{t-i}y_{t-\tau}) = \gamma_{\tau-i}$, equation (24) gives

$$\sum_i \alpha_i \gamma_{\tau-i} = \begin{cases} \sigma_\varepsilon^2, & \text{if } \tau = 0; \\ 0, & \text{if } \tau > 0. \end{cases} \tag{26}$$

The second of these is a homogeneous difference equation which enables us to generate the sequence $\{\gamma_p, \gamma_{p+1}, \ldots\}$ once $p$ starting values $\gamma_0, \gamma_1, \ldots, \gamma_{p-1}$ are known. By letting $\tau = 0, 1, \ldots, p$ in (26), we generate a set of $p+1$ equations which can be arrayed in matrix form as follows:

$$\begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_p \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_p & \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{bmatrix} \begin{bmatrix} 1 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} \sigma_\varepsilon^2 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{27}$$

These are called the Yule–Walker equations, and they can be used either for generating the values $\gamma_0, \gamma_1, \ldots, \gamma_p$ from the values $\alpha_1, \ldots, \alpha_p, \sigma_\varepsilon^2$ or vice versa.

**Example.** To illustrate the two uses of the Yule–Walker equations, let us consider the second-order autoregressive process. In that case, we have

$$\begin{aligned} \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_1 & \gamma_0 & \gamma_1 \\ \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} &= \begin{bmatrix} \alpha_2 & \alpha_1 & \alpha_0 & 0 & 0 \\ 0 & \alpha_2 & \alpha_1 & \alpha_0 & 0 \\ 0 & 0 & \alpha_2 & \alpha_1 & \alpha_0 \end{bmatrix} \begin{bmatrix} \gamma_2 \\ \gamma_1 \\ \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} \\ &= \begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 \\ \alpha_1 & \alpha_0 + \alpha_2 & 0 \\ \alpha_2 & \alpha_1 & \alpha_0 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} \sigma_\varepsilon^2 \\ 0 \\ 0 \end{bmatrix}. \end{aligned} \tag{28}$$

Given $\alpha_0 = 1$ and the values for $\gamma_0, \gamma_1, \gamma_2$, we can find $\sigma_\varepsilon^2$ and $\alpha_1, \alpha_2$. Conversely, given $\alpha_0, \alpha_1, \alpha_2$ and $\sigma_\varepsilon^2$, we can find $\gamma_0, \gamma_1, \gamma_2$. It is worth recalling at this juncture that the normalisation $\sigma_\varepsilon^2 = 1$ might have been chosen instead of $\alpha_0 = 1$. This would have rendered the equations more easily intelligible. Notice also how the matrix following the first equality is folded across the axis which divides it vertically to give the matrix which follows the second equality. Pleasing effects of this sort often arise in time-series analysis.

**Autoregressive Moving Average Processes**

The autoregressive moving-average process of orders $p$ and $q$, which is referred to as the ARMA$(p, q)$ process, is defined by the equation

$$\begin{aligned}\alpha_0 y(t) + \alpha_1 y(t-1) + \cdots + \alpha_p y(t-p) \\ = \mu_0 \varepsilon(t) + \mu_1 \varepsilon(t-1) + \cdots + \mu_q \varepsilon(t-q).\end{aligned} \tag{29}$$

The equation is normalised by setting $\alpha_0 = 1$ and by setting either $\mu_0 = 1$ or $\sigma_\varepsilon^2 = 1$. A more summary expression for the equation is $\alpha(L)y(t) = \mu(L)\varepsilon(t)$. Provided that the roots of the equation $\alpha(z) = 0$ lie outside the unit circle, the process can be represented by the equation $y(t) = \alpha^{-1}(L)\mu(L)\varepsilon(t)$ which corresponds to an infinite-order moving-average process. Conversely, provided the roots of the equation $\mu(z) = 0$ lie outside the unit circle, the process can be represented by the equation $\mu^{-1}(L)\alpha(L)y(t) = \varepsilon(t)$ which corresponds to an infinite-order autoregressive process.

By considering the moving-average form of the process, and by noting the form of the autocovariance generating function for such a process which is given by equation (16), it can be seen that the autocovariance generating function for the autoregressive moving-average process is

$$\gamma(z) = \sigma_\varepsilon^2 \frac{\mu(z)\mu(z^{-1})}{\alpha(z)\alpha(z^{-1})}. \tag{30}$$

This generating function, which is of some theoretical interest, does not provide a practical means of finding the autocovariances. To find these, let us consider multiplying the equation $\sum_i \alpha_i y_{t-i} = \sum_i \mu_i \varepsilon_{t-i}$ by $y_{t-\tau}$ and taking expectations. This gives

$$\sum_i \alpha_i \gamma_{i-\tau} = \sum_i \mu_i \delta_{i-\tau}, \tag{31}$$

where $\gamma_{i-\tau} = E(y_{t-i}y_{t-\tau})$ and $\delta_{i-\tau} = E(\varepsilon_{t-i}y_{t-\tau})$. Since $\varepsilon_{t-i}$ is uncorrelated with $y_{t-\tau}$ whenever it is subsequent to the latter, it follows that $\delta_{i-\tau} = 0$ if

$\tau > i$. Since the index $i$ in the RHS of the equation (31) runs from 0 to $q$, it follows that

$$\sum_i \alpha_i \gamma_{i-\tau} = 0 \quad \text{if} \quad \tau > q. \tag{32}$$

Given the $q+1$ nonzero values $\delta_0, \delta_1, \ldots, \delta_q$, and $p$ initial values $\gamma_0, \gamma_1, \ldots, \gamma_{p-1}$ for the autocovariances, the equations can be solved recursively to obtain the subsequent values $\{\gamma_p, \gamma_{p+1}, \ldots\}$.

To find the requisite values $\delta_0, \delta_1, \ldots, \delta_q$, consider multiplying the equation $\sum_i \alpha_i y_{t-i} = \sum_i \mu_i \varepsilon_{t-i}$ by $\varepsilon_{t-\tau}$ and taking expectations. This gives

$$\sum_i \alpha_i \delta_{\tau-i} = \mu_\tau \sigma_\varepsilon^2, \tag{33}$$

where $\delta_{\tau-i} = E(y_{t-i}\varepsilon_{t-\tau})$. The equation may be rewritten as

$$\delta_\tau = \frac{1}{\alpha_0}\Big(\mu_\tau \sigma_\varepsilon^2 - \sum_{i=1} \delta_{\tau-i}\Big), \tag{34}$$

and, by setting $\tau = 0, 1, \ldots, q$, we can generate recursively the required values $\delta_0, \delta_1, \ldots, \delta_q$.

**Example.** Consider the ARMA(2, 2) model which gives the equation

$$\alpha_0 y_t + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} = \mu_0 \varepsilon_t + \mu_1 \varepsilon_{t-1} + \mu_2 \varepsilon_{t-2}. \tag{35}$$

Multiplying by $y_t$, $y_{t-1}$ and $y_{t-2}$ and taking expectations gives

$$\begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_1 & \gamma_0 & \gamma_1 \\ \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \delta_0 & \delta_1 & \delta_2 \\ 0 & \delta_0 & \delta_1 \\ 0 & 0 & \delta_0 \end{bmatrix} \begin{bmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \end{bmatrix}. \tag{36}$$

Multiplying by $\varepsilon_t$, $\varepsilon_{t-1}$ and $\varepsilon_{t-2}$ and taking expectations gives

$$\begin{bmatrix} \delta_0 & 0 & 0 \\ \delta_1 & \delta_0 & 0 \\ \delta_2 & \delta_1 & \delta_0 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \sigma_\varepsilon^2 & 0 & 0 \\ 0 & \sigma_\varepsilon^2 & 0 \\ 0 & 0 & \sigma_\varepsilon^2 \end{bmatrix} \begin{bmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \end{bmatrix}. \tag{37}$$

When the latter equations are written as

$$\begin{bmatrix} \alpha_0 & 0 & 0 \\ \alpha_1 & \alpha_0 & 0 \\ \alpha_2 & \alpha_1 & \alpha_0 \end{bmatrix} \begin{bmatrix} \delta_0 \\ \delta_1 \\ \delta_2 \end{bmatrix} = \sigma_\varepsilon^2 \begin{bmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \end{bmatrix}, \tag{38}$$

they can be solved recursively for $\delta_0$, $\delta_1$ and $\delta_2$ on the assumption that that the values of $\alpha_0$, $\alpha_1$, $\alpha_2$ and $\sigma_\varepsilon^2$ are known. Notice that, when we adopt the normalisation $\alpha_0 = \mu_0 = 1$, we get $\delta_0 = \sigma_\varepsilon^2$. When the equations (36) are rewritten as

$$\begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 \\ \alpha_1 & \alpha_0 + \alpha_2 & 0 \\ \alpha_2 & \alpha_1 & \alpha_0 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & 0 \\ \mu_2 & 0 & 0 \end{bmatrix} \begin{bmatrix} \delta_0 \\ \delta_1 \\ \delta_2 \end{bmatrix}, \tag{39}$$

they can be solved for $\gamma_0$, $\gamma_1$ and $\gamma_2$. Thus the starting values are obtained which enable the equation

$$\alpha_0\gamma_\tau + \alpha_1\gamma_{\tau-1} + \alpha_2\gamma_{\tau-2} = 0; \quad \tau > 2 \tag{40}$$

to be solved recursively to generate the succeeding values $\{\gamma_3, \gamma_4, \ldots\}$ of the autocovariances.

**Minimum Mean-Square Error Prediction**

Imagine that $y(t)$ is a stationary stochastic process with $E\{y(t)\} = 0$. We may be interested in predicting values of this process several periods into the future on the basis of its observed history. This history is contained in our so-called information set. In practice, the latter is always a finite set $\{y_t, y_{t-1}, \ldots, y_{t-p}\}$ representing the recent past. Nevertheless, in developing the theory of prediction, it is also useful to consider an infinite information set $\{y_t, y_{t-1}, \ldots, y_{t-p}, \ldots\}$ representing the entire past.

We shall denote the prediction of $y_{t+m}$ which is made at the time $t$ by $\hat{y}_{t+m|t}$ or by $\hat{y}_{t+m}$ when it is clear that we a predicting $m$ steps ahead.

The criterion by which we usually judge the performance of an estimator or predictor $\hat{y}$ of a random variable $y$ is its mean-square error defined by $E\{(y - \hat{y})^2\}$. If all of the available information on $y$ is summarised in its marginal distribution, then the minimum mean-square error prediction is simply the expected value $E(y)$. However, if $y$ is statistically related to another random variable $x$ whose value we can observe, and if we know the form of the joint distribution of $x$ and $y$, then the minimum mean-square error prediction of $y$ is the conditional expectation $E(y|x)$. We may state this proposition formally:

(41) Let $\hat{y} = \hat{y}(x)$ be the conditional expectation of $y$ given $x$ which is also expressed as $\hat{y} = E(y|x)$. Then we have $E\{(y - \hat{y})^2\} \leq E\{(y - \pi)^2\}$, where $\pi = \pi(x)$ is any other function of $x$.

*Proof.* Consider

$$\begin{aligned} E\{(y - \pi)^2\} &= E[\{(y - \hat{y}) + (\hat{y} - \pi)\}^2] \\ &= E\{(y - \hat{y})^2\} + 2E\{(y - \hat{y})(\hat{y} - \pi)\} + E\{(\hat{y} - \pi)^2\} \end{aligned} \tag{42}$$

In the second term, we have

$$\begin{aligned} E\{(y-\hat{y})(\hat{y}-\pi)\} &= \int_x \int_y (y-\hat{y})(\hat{y}-\pi) f(x,y)\partial y \partial x \\ &= \int_x \left\{ \int_y (y-\hat{y}) f(y|x)\partial y \right\} (\hat{y}-\pi) f(x)\partial x \\ &= 0. \end{aligned} \tag{43}$$

Here the second equality depends upon the factorisation $f(x,y) = f(y|x)f(x)$ which expresses the joint probability density function of $x$ and $y$ as the product of the conditional density function of $y$ given $x$ and the marginal density function of $x$. The final equality depends upon the fact that $\int (y-\hat{y})f(y|x)\partial y = E(y|x) - E(y|x) = 0$. Therefore $E\{(y-\pi)^2\} = E\{(y-\hat{y})^2\} + E\{(\hat{y}-\pi)^2\} \geq E\{(y-\hat{y})^2\}$ and our assertion is proved.

We might note that the definition of the conditional expectation implies that

$$\begin{aligned} E(xy) &= \int_x \int_y xy f(x,y)\partial y \partial x \\ &= \int_x x \left\{ \int_y y f(y|x)\partial y \right\} f(x)\partial x \\ &= E(x\hat{y}). \end{aligned} \tag{44}$$

When the equation $E(xy) = E(x\hat{y})$ is rewritten as

$$E\{x(y-\hat{y})\} = 0, \tag{45}$$

it may be described as an orthogonality condition. This condition indicates that the prediction error $y - \hat{y}$ is uncorrelated with $x$. The result is intuitively appealing; for, if the error were correlated with $x$, we should not using the information of $x$ efficiently in forming $\hat{y}$.

The proposition of (41) is readily generalised to accommodate the case where, in place of the scalar $x$, we have a vector $x = [x_1, \ldots, x_p]'$. This generalisation indicates that the minimum-mean-square-error prediction of $y_{t+m}$ given the information in $\{y_t, y_{t-1}, \ldots, y_{t-p}\}$ is the conditional expectation $E(y_{t+m}|y_t, y_{t-1}, \ldots, y_{t-p})$.

In order to determine the conditional expectation of $y_{t+m}$ given $\{y_t, y_{t-1}, \ldots, y_{t-p}\}$, we need to known the functional form of the joint probability density function all of these variables. In lieu of precise knowledge, we are often prepared to assume that the distribution is normal. In that case, it follows that the conditional expectation of $y_{t+m}$ is a linear function of $\{y_t, y_{t-1}, \ldots, y_{t-p}\}$;

and so the problem of predicting $y_{t+m}$ becomes a matter of forming a linear regression. Even if we are not prepared to assume that the joint distribution of the variables in normal, we may be prepared, nevertheless, to base our prediction of $y$ upon a linear function of $\{y_t, y_{t-1}, \ldots, y_{t-p}\}$. In that case, we satisfy the criterion of minimum mean-square error linear prediction by forming $\hat{y}_{t+m} = \sum \phi_j y_{t-j+1}$ from the values $\phi_1, \ldots, \phi_{p+1}$ which minimise

$$\begin{aligned} E\left\{(y_{t+m} - \hat{y}_{t+m})^2\right\} &= E\left\{\left(y_{t+m} - \sum_{j=1}^{p+1} \phi_j y_{t-j+1}\right)^2\right\} \\ &= \gamma_0 - 2\sum_j \phi_j \gamma_{m+j-1} + \sum_i \sum_j \phi_i \phi_j \gamma_{i-j}. \end{aligned} \tag{46}$$

This is a linear least-squares regression problem which leads to a set of $p+1$ orthogonality conditions described as the normal equations:

$$\begin{aligned} E\{(y_{t+m} - \hat{y}_{t+m}) y_{t-j+1}\} &= \gamma_{m+j-1} - \sum_{i=1}^{p} \phi_i \gamma_{i-j} \\ &= 0 \quad ; \quad j = 1, \ldots, p+1. \end{aligned} \tag{47}$$

In matrix terms, we have

$$\begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_p \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_p & \gamma_{p-1} & \cdots & \gamma_0 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{p+1} \end{bmatrix} = \begin{bmatrix} \gamma_m \\ \gamma_{m+1} \\ \vdots \\ \gamma_{m+p} \end{bmatrix}. \tag{48}$$

Notice that, for the one-step-ahead prediction of $y_{t+1}$, these are nothing but the Yule–Walker equations.

**Forecasting with ARMA Models**

So far, we have avoided making any specific assumptions about the nature of the process $y(t)$ other than that it can be represented by an infinite-order moving average. We are greatly assisted in the business of developing practical forecasting procedures if we can assume that $y(t)$ is generated by an ARMA process such that

$$y(t) = \frac{\mu(L)}{\alpha(L)}\varepsilon(t) = \psi(L)\varepsilon(t). \tag{49}$$

We shall continue to assume, for the sake of simplicity, that the forecasts are based on the information contained in the infinite set $\{y_t, y_{t-1}, \ldots, y_{t-p}, \ldots\}$

comprising all values that have been taken by the variable up to the present time $t$. Knowing the parameters in $\psi(L)$ enables us to recover the sequence $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$ from the sequence $\{y_t, y_{t-1}, y_{t-2}, \ldots\}$ and vice versa; so either of these can be regarded as our information set.

Let us write the realisations of equation (49) as

$$y_{t+m} = \sum_{i=0}^{m-1} \psi_i \varepsilon_{t+m-i} + \sum_{i=m}^{\infty} \psi_i \varepsilon_{t+m-i}. \tag{50}$$

Here the first term on the RHS embodies disturbances subsequent to the time $t$ when the forecast is made, and the second term embodies disturbances which are within the information set $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$. Let us now define a forecasting function, based on the information set, which takes the form of

$$\hat{y}_{t+m} = \sum_{i=m}^{\infty} \rho_i \varepsilon_{t+m-i}. \tag{51}$$

Then, given that $\varepsilon(t)$ is a white-noise process, it follows that the mean square of the error in the forecast $m$ periods ahead is given by

$$E\{(y_{t+m} - \hat{y}_{t+m})^2\} = \sigma_\varepsilon^2 \sum_{i=0}^{m-1} \psi_i^2 + \sigma_\varepsilon^2 \sum_{i=m}^{\infty} (\psi_i - \rho_i)^2. \tag{52}$$

Clearly, the mean-square error is minimised by setting $\rho_i = \psi_i$; and so the optimal forecast is given by

$$\hat{y}_{t+m} = \sum_{i=m}^{\infty} \psi_i \varepsilon_{t+m-i}. \tag{53}$$

This might have been derived from the the equation $y(t+m) = \psi(L)\varepsilon(t+m)$, which generates the the true value of $y_{t+m}$, simply by putting zeros in place of the unobserved disturbances $\varepsilon_{t+1}, \varepsilon_{t+2}, \ldots, \varepsilon_{t+m}$ which lie in the future when the forecast is made. Notice that, as the lead time $m$ of the forecast increases, the mean-square error of the forecast tends to the value of

$$V\{y(t)\} = \sigma_\varepsilon^2 \sum \psi_i^2 \tag{54}$$

which is nothing but the variance of the process $y(t)$.

We can also derive the optimal forecast of (45) by specifying that the forecast error should be uncorrelated with the disturbances up to the time of making the forecast. For, if the the forecast errors were correlated with some of the elements of our information set, then, as we have noted before, we would

not be using the information efficiently, and we could not be generating optimal forecasts. To demonstrate this result anew, let us consider the covariance between the forecast error and the disturbance $\varepsilon_{t-i}$:

$$\begin{aligned} E\{(y_{t+m} - \hat{y}_{t+m})\varepsilon_{t-i}\} &= \sum_{k=1}^{m} \psi_{m-k} E(\varepsilon_{t+k}\varepsilon_{t-i}) \\ &\quad + \sum_{j=0}^{\infty} (\psi_{m+j} - \rho_{m+j}) E(\varepsilon_{t-j}\varepsilon_{t-i}) \\ &= \sigma_{\varepsilon}^2 (\psi_{m+i} - \rho_{m+i}). \end{aligned} \tag{55}$$

Here the final equality follows from the fact that

$$E(\varepsilon_{t-j}\varepsilon_{t-i}) = \begin{cases} \sigma_{\varepsilon}^2, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases} \tag{56}$$

If the covariance in (55) is to be equal to zero for all values of $i \geq 0$, then we must have $\rho_i = \psi_i$ for all $i$, which means that our forecasting function must be the one that we have already specified under (53).

It is helpful, sometimes, to have a functional notation for describing the process which generates the $m$-steps-ahead forecast. The notation provided by Whittle (1963) is widely used. To derive this, let us begin by writing

$$y(t+m|t) = \left\{L^{-m}\psi(L)\right\}\varepsilon(t). \tag{57}$$

On the LHS, we have not only the lagged sequences $\varepsilon(t), \varepsilon(t-1), \ldots$ but also the sequences $\varepsilon(t+m) = L^{-m}\varepsilon(t), \ldots, \varepsilon(t+1) = L^{-1}\varepsilon(t)$, all of which are associated with negative powers of $L$. Let $\{L^{-m}\psi(L)\}_+$ be defined as the part of the operator containing only positive powers of $L$. Then we can express the forecasting function as

$$\begin{aligned} \hat{y}(t+m|t) &= \{L^{-m}\psi(L)\}_+ \varepsilon(t) \\ &= \left\{\frac{\psi(L)}{L^m}\right\}_+ \frac{1}{\psi(L)} y(t). \end{aligned} \tag{58}$$

**The Forecasts as Conditional Expectations**

We have already seen that we can regard the optimal (minimum mean-square error) forecast of $y_{t+m}$ as the conditional expectation of $y_{t+m}$ given the values of $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$ or $\{y_t, y_{t-1}, y_{t-2}, \ldots\}$. Let us denote the forecast by $\hat{y}_{t+m} = E_t(y_{t+m})$ where the subscript on the operator is to indicate that

the expectation is conditional upon the information available at time $t$. On applying the operator to the sequences $y(t)$ and $\varepsilon(t)$, we find that

$$\begin{aligned} E_t(y_{t+k}) &= \hat{y}_{t+k} \quad ; \quad k > 0 \\ E_t(y_{t-j}) &= y_{t-j} \quad ; \quad j \geq 0 \\ E_t(\varepsilon_{t+k}) &= 0 \quad ; \quad k > 0 \\ E_t(\varepsilon_{t-j}) &= \varepsilon_{t-j} \quad ; \quad j \geq 0. \end{aligned} \tag{59}$$

In this notation, the forecast $m$ periods ahead is

$$\begin{aligned} E_t(y_{t+m}) &= \sum_{k=1}^{m} \psi_{m-k} E_t(\varepsilon_{t+k}) + \sum_{j=0}^{\infty} \psi_{m+j} E_t(\varepsilon_{t-j}) \\ &= \sum_{j=0}^{\infty} \psi_{m+j} \varepsilon_{t-j}. \end{aligned} \tag{60}$$

In practice, we may generate the forecasts using a recursion based on the equation

$$\begin{aligned} y(t) = &-\{\alpha_1 y(t-1) + \alpha_2 y(t-2) + \cdots + \alpha_p y(t-p)\} \\ &+ \mu_0 \varepsilon(t) + \mu_1 \varepsilon(t-1) + \cdots + \mu_q \varepsilon(t-q). \end{aligned} \tag{61}$$

By taking the conditional expectation of this function, we get

$$\begin{aligned} \hat{y}_{t+m} = &-\{\alpha_1 \hat{y}_{t+m-1} + \cdots + \alpha_p y_{t+m-p}\} \\ &+ \mu_m \varepsilon_t + \cdots + m u_q \varepsilon_{t+m-q} \quad \text{when} \quad 0 < m \leq p, q, \end{aligned} \tag{62}$$

$$\hat{y}_{t+m} = -\{\alpha_1 \hat{y}_{t+m-1} + \cdots + \alpha_p y_{t+m-p}\} \quad \text{if} \quad q < m \leq p, \tag{63}$$

$$\begin{aligned} \hat{y}_{t+m} = &-\{\alpha_1 \hat{y}_{t+m-1} + \cdots + \alpha_p \hat{y}_{t+m-p}\} \\ &+ \mu_m \varepsilon_t + \cdots + \mu_q \varepsilon_{t+m-q} \quad \text{if} \quad p < m \leq q, \end{aligned} \tag{64}$$

and

$$\hat{y}_{t+m} = -\{\alpha_1 \hat{y}_{t+m-1} + \cdots + \alpha_p \hat{y}_{t+m-p}\} \quad \text{when} \quad p, q < m. \tag{65}$$

We can see from (65) that, for $m > p, q$, the forecasting function becomes a $p$th-order homogeneous difference equation in $y$. The $p$ values of $y(t)$ from $t = r = \max(p, q)$ to $t = r - p + 1$ serve as the starting values for the equation.

The behaviour of the forecast function beyond the reach of the starting values can be characterised in terms of the roots of the autoregressive operator. We can assume that none of the roots of $\alpha(L) = 0$ lie inside the unit circle. If all of the roots are less than unity, then $\hat{y}_{t+m}$ will converge to zero as $m$ increases. If one of the roots of $\alpha(L) = 0$ is unity, then we have and ARIMA$(p, 1, q)$ model; and the general solution of the homogeneous equation of (65) will include a constant term which represents the product of the unit root with an coefficient which is determined by the starting values. Hence the the forecast will tend to a nonzero constant. If two of the roots are unity, then the the general solution will embody a linear time trend which is the asymptote to which the forecasts will tend. In general, if $d$ of the roots are unity, then the general solution will comprise a polynomial in $t$ of order $d - 1$.

The forecasts can be updated easily once the coefficients in the expansion of $\psi(L) = \mu(L)/\alpha(L)$ have been obtained. Consider

$$\begin{aligned} \hat{y}_{(t+1)+m} &= \{\psi_m\varepsilon_{t+1} + \psi_{m+1}\varepsilon_t + \psi_{m+2}\varepsilon_{t-1} + \cdots\} \quad \text{and} \\ \hat{y}_{t+(m+1)} &= \{\psi_{m+1}\varepsilon_t + \psi_{m+2}\varepsilon_{t-1} + \cdots\}. \end{aligned} \tag{66}$$

The first of these is the forecast for $m$ periods ahead made at time $t+1$ whilst the second is the forecast for $m+1$ periods ahead made at time $t$. We can easily see that

$$\hat{y}_{(t+1)+m} = \hat{y}_{t+(m+1)} + \psi_m\varepsilon_{t+1}, \tag{67}$$

where $\varepsilon_{t+1} = \hat{y}_{t+1} - y_{t+1}$ is the current disturbance at time $t+1$. The later is also the prediction error of the one-step-ahead forecast made at time $t$.

**Example.** Consider the AR(41) process. We have

$$y(t+m) = \phi y(t+m-1) + \varepsilon(t+m). \tag{68}$$

On applying the operator $E_t$ to this equation we obtain the following:

$$\begin{aligned} \hat{y}(t+1) &= \phi y(t) \\ \hat{y}(t+2) &= \phi\hat{y}(t+1) = \phi^2 y(t) \\ &\vdots \\ \hat{y}(t+m) &= \phi\hat{y}(t+m-1) = \phi^m y(t). \end{aligned} \tag{69}$$

Given that $y(t) = \varepsilon(t)/(1-\phi L) = \{\varepsilon(t) + \phi\varepsilon(t-1) + \phi^2\varepsilon(t-2) + \cdots\}$, it follows from (44) that

$$E\{(y_{t+m} - \hat{y}_{t+m})^2\} = \sigma_\varepsilon^2\left\{1 + \phi^2 + \phi^4 + \cdots + \phi^{2(m-1)}\right\}. \tag{70}$$

As $m \to \infty$, this tends to $\sigma_\varepsilon^2/(1-\phi^2)$ which is just the variance of the AR(41) process.

**Example.** *(Exponential Smoothing).* A common forecasting procedure is exponential smoothing. This depends upon taking a weighted average of past values of the time series with the weights following a geometrically declining pattern. The function generating the one-step-ahead forecasts can be written as

$$\begin{aligned}\hat{y}(t+1) &= \frac{(1-\theta)}{1-\theta L}y(t) \\ &= (1-\theta)\left\{y(t) + \theta y(t-1) + \theta^2 y(t+2) + \cdots\right\}.\end{aligned} \tag{71}$$

On multiplying both sides of this equation by $1-\theta L$ and rearranging, we get

$$\hat{y}(t+1) = \theta\hat{y}(t) + (1-\theta)y(t), \tag{72}$$

which shows that the current forecast for one step ahead is a convex combination of the previous forecast and the value that actually transpired.

It is possible to show that the method of exponential smoothing corresponds to the optimal forecasting procedure for the ARIMA$(0,1,1)$ model $(1-L)y(t) = (1-\theta L)\varepsilon(t)$. To see this, let us consider the ARMA$(1,1)$ model $y(t) - \phi y(t-1) = \varepsilon(t) - \theta\varepsilon(t-1)$. This gives

$$\begin{aligned}\hat{y}(t+1) &= \phi y(t) - \theta\varepsilon(t) \\ &= \phi y(t) - \theta\frac{(1-\phi L)}{1-\theta L}y(t) \\ &= \frac{\{(1-\theta L)\phi - (1-\phi L)\theta\}}{1-\theta L}y(t) \\ &= \frac{(\phi-\theta)}{1-\theta L}y(t)\end{aligned} \tag{73}$$

On setting $\phi = 1$, which converts the ARMA$(1,1)$ model to an ARIMA$(0,1,1)$ model, we obtain precisely the forecasting function of (60).