Division of Research
Graduate School of Business Administration
The University of Michigan

August 1974

ANALYZING AND FORECASTING TIME SERIES
Part 1: Methodology

Working Paper No. 91
by

W. Allen Spivey
William J. Wrobleski

The University of Michigan

TABLE OF CONTENTS

1. INTRODUCTION

We take the point of view that analyzing and forecasting a time series are interrelated activities. One uses time series analysis as an important model building strategy whereby intuitive ideas about the underlying time series data are translated into an appropriate forecasting model.

Two major approaches to time series analysis are available. One, which we will call components analysis, regards the time series as being composed of several influences or components which are generally taken to be trend and cyclical, seasonal, and irregular or random movements. Oftentimes in this approach, trend and seasonal influences are modeled in a deterministic manner; trend may be regarded as a polynomial of a given degree and the seasonal component may be modeled by a trigonometric function with given period and amplitude. Random influences are usually assumed to have a simple probability structure, i.e., they are treated as independent, identically distributed random variables having zero mean and finite variance. Because of the simple role of probability in this approach, estimation of the trend and cyclical component, as well as the seasonal component, is carried out using fairly simple filtering procedures such as differencing, forming moving averages, considering

ratios involving moving averages, and others rather than by using parametric statistical estimation procedures.

The other major approach is to regard a time series as an observed sample function representing a realization of an underlying stochastic process. A stochastic process can be stationary or non-stationary; it can be a moving average process, an autoregressive process, or one which involves both autoregressive and moving average terms. Moreover, associated with a stochastic process are the mean and autocovariance functions, the autocorrelation function, and its power spectrum. When two or more processes are considered, their cross-correlation functions, cospectra, and other characteristics may be examined. This injects a full array of probabilistic concepts and methods into time series analysis and forecasting. The more elaborate probability framework permits a more complete consideration of statistical estimation than is generally considered in components analysis.

2. COMPONENTS TIME SERIES ANALYSIS

The components of a time series can be assumed to be combined additively, multiplicatively, or some mixture of the two. Thus we can speak of additive, multiplicative, or mixed models. Specifically, if TC(t) indicates the trend-cyclical component, S(t) the

seasonal, and $e(t)$ the random component, an additive model is given by

$$Y(t) = TC(t) + S(t) + e(t), \tag{2.1}$$

where $Y(t)$ denotes the observed data. A multiplicative model is

$$Y(t) = TC(t)S(t)e(t), \tag{2.2}$$

and an example of a mixed model is

$$Y(t) = TC(t)S(t) + e(t). \tag{2.3}$$

For purposes of forecasting with a components model, one approach is to estimate the components of $Y(t)$ appropriately, and use these estimates to obtain a forecast $\hat{Y}(t+\tau)$ of the expected value of $Y(t)$ τ periods ahead as follows.[1] If $\hat{TC}(t)$ and $\hat{S}(t)$ are estimates of $TC(t)$ and $S(t)$, respectively, then the τ periods ahead forecast of $Y(t)$ is taken as

$$\hat{Y}(t+\tau) = \hat{TC}(t+\tau) + \hat{S}(t+\tau) \tag{2.4}$$

in the additive model or as

$$\hat{Y}(t+\tau) = \hat{TC}(t+\tau)\hat{S}(t+\tau) \tag{2.5}$$

for either the multiplicative or mixed model.

---

[1] The forecast of the expected value of $Y(t+\tau)$ is often called the forecast value of $Y(t)$.

To determine prediction intervals for $\hat{Y}(t+\tau)$ a study of the random component $\hat{e}(t)$ would be required. For example, if the random component were considered to be a disturbance term which from period to period could be assumed to be independent and identically distributed with zero mean and unknown variance, an estimate of this variance as well as $\hat{TC}(t)$ and $\hat{S}(t)$ would be necessary.

Methods of estimating components vary from simple ad hoc specification to elaborate computer-based manipulations of the data. Detrending methods illustrate this variety very well. Among the simpler methods is the calculation of first or higher-order differences of the data. The order of differencing used corresponds to one's perception of the trend as represented by a polynomial of given degree, and one might use the variate difference method to estimate the degree of the polynomial representing the trend and cyclical component. Other detrending procedures involve using a moving average as well as polynomial regression.

A wide variety of deseasonalizing procedures is also available. These range from manipulation of moving averages, such as the Census II method and its variants, to regression analysis using dummy variables, to more sophisticated harmonic and spectral analyses.

Developing suitable estimates of these components is a delicate problem in time series analysis and these estimates obviously

influence whatever forecasting model is developed subsequently. Unfortunately, from a forecasting point of view, the methods of statistical estimation of these components are inadequate at the present time. Rigorous definitions of trend and seasonal influences have not been developed. Yet another difficulty is that some widely used procedures for separating one component from a time series have been shown to separate parts of others as well. Moreover, a circularity in estimation is present in some procedures. For example, if one wishes to estimate the TC(t) component, a standard method is to estimate the seasonal component and deasonalize the data, producing a time series consisting of TC(t) and $\check{e}(t)$. Unfortunately, in order to estimate S(t) one must make an initial assumption concerning TC(t). Therefore some implicit detrending of the data is present in the process of attempting to estimate TC(t), thus producing an evident circularity in the procedure.

Despite important shortcomings, which include lack of rigorous definitions for TC(t) and S(t) and the fact that many unresolved inference problems remain, components analysis continues to occupy an important place in time series analysis. The literature is large and for over 70 years has received wide attention in both theoretical and applied fields. Our survey is largely directed to time series analysis in a probabilistic framework and we do not give a full treatment

of components analysis here. However, as the reader will see, one is not entirely free of some form of components analysis and the brief discussion above is useful in the material that follows.

3. TIME SERIES DATA AS A STOCHASTIC PROCESS

A stochastic process can be viewed as a family of random variables $\{X(\omega, t)\}$ defined on an appropriate probability space $(\Omega, \mathcal{F}, P)$ where $\omega \varepsilon \Omega$ and the index t belongs to a set T. Here Ω denotes a sample space, $\mathcal{F}$ is a σ-field of events associated with Ω, and P is a probability measure.

For all purposes in this paper, the set T is either an infinite subset of the set of integers or is an interval on the real line. Whenever T is a countable set we say that the stochastic process is a discrete parameter process; otherwise we say it is a continuous parameter process.

For an alternate view, suppose that ω is a fixed point in Ω. If we treat the corresponding values of the random variable $X(\omega, t)$ for the fixed ω as a function of t, then we arrive at the concept of a sample function of the stochastic process $\{X(\omega, t)\}$. We suppress explicit reference to ω in this case, simply using X(t) to denote the sample function.

A realized (observed) sample function is called a time series. Thus by means of this approach and through appropriate statistical

analysis and estimation applied to time series data, it may be possible to make inferences concerning important properties of the stochastic process generating the data. These properties also provide a basis for the development of forecasting models, as will be discussed in later sections.

4. CLASSIFICATION OF STOCHASTIC PROCESSES AND RELATED CONCEPTS

Associated with either of the points of view in the previous section are the families of finite dimensional distribution functions,

$$F(x_1, \cdots, x_n; t_1, \cdots, t_n) = P(X(t_1) \leqq x_1, \cdots, X(t_n) \leqq x_n)$$

where $\{t_1, \cdots, t_n\}$ is any finite subset of the index set T. If it turns out that these families of distribution functions satisfy the condition that

$$(4.1) \quad F(x_1, \cdots, x_n; t_1, \cdots, t_n) = F(x_1, \cdots, x_n; t_1+h, \cdots, t_n+h)$$

for any displacement h of $t_1, t_2, \cdots, t_n, t$, then the stochastic process is said to be a strictly stationary stochastic process.

Relating this important concept to the sample function interpretation, we observe that if a time series displays a pronounced trend that persists over a long period of time, the distribution functions

could not satisfy the condition (4. 1) and one would in this case say that the underlying stochastic process is not strictly stationary.

Using the family of distribution functions introduced above, we define the mean function of a stochastic process as

$$M(t) = \int_{-\infty}^{\infty} x\,dF(x;t) = E[X(t)]. \tag{4.2}$$

The variance function is

$$V(t) = \int_{-\infty}^{\infty} (x(t) - M(t))^2 dF(x;t) = E[(X(t) - M(t))^2] \tag{4.3}$$

and the covariance function between $X(t_1)$ and $X(t_2)$ is given by

$$C(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_1 - M(t_1))(x_2 - M(t_2))\,dF(x_1, x_2; t_1, t_2) \tag{4.4}$$

$$= E[X(t_1) - M(t_1))(X(t_2) - M(t_2))].$$

Note that when $t_1 = t_2 = t$ we get $C(t, t) = V(t)$.

The correlation function between $X(t_1)$ and $X(t_2)$ is

$$\rho(t_1, t_2) = \frac{C(t_1, t_2)}{\sqrt{V(t_1)V(t_2)}}. \tag{4.5}$$

It is obvious that when a stochastic process is strictly stationary we have $M(t) = M(t+h)$ and $V(t) = V(t+h)$, since $F(x;t) = F(x;t+h)$. Thus the mean $M(t)$ and the variance $V(t)$ of a strictly stationary process must be constant and cannot depend on t. In addition,

$$C(t_1, t_2) = C(t_1+h, t_2+h)$$

and

$$\rho(t_1, t_2) = \rho(t_1+h, t_2+h)$$

since $F(x_1, x_2; t_1, t_2) = F(x_1, x_2; t_1+h, t_2+h)$. Furthermore, taking $t_1 = t-h$ and $t_2 = t$ gives

$$C(t_1, t_2) = C(t-h, t) = C(t, t+h) = C(h) \tag{4.6}$$

and we see that for strictly stationary processes the covariance is a function of the displacement h alone. Moreover, if we let $h = 0$ in (4.6) we get

$$C(0) = C(t, t) = V(t) = V(0),$$

since in the strictly stationary case V(t) is a constant. A similar argument shows that

$$\rho(t_1, t_2) = \rho(t-h, t) = \rho(t, t+h) = \rho(h).$$

In this case the functions C(h) and $\rho(h)$, regarded as functions of h, are called the autocovariance and autocorrelation functions, respectively.

An important class of stochastic processes, less restricted than the class of strictly stationary processes, is that of covariance

stationary processes. These and other still less restricted classes are oftentimes of interest in time series analysis.

A stochastic process is said to be stationary in the mean if the mean function $M(t) = E[X(t)]$ is a constant (and does not depend on t). A process is said to be covariance stationary (or stationary in the wide sense) if

$$C(t, t+h) = E[(X(t) - M(t))(X(t+h) - M(t+h))]$$

is a function of the displacement h only.

This condition in turn implies that $M(t) = E[X(t)]$ and $V(t) = E[(X(t) - M(t))^2]$ are constants, neither depending on t. In other words, covariance stationarity implies stationarity in the mean and stationarity in the variance.

5. POWER SPECTRA AND COVARIANCE STATIONARY PROCESSES

We have seen that when a stochastic process is covariance stationary its autocovariance function depends only on the displacement of t. The covariance function C(k) of a discrete parameter covariance stationary process regarded as a function of the displacement k is a positive semi-definite function. Thus it can be represented as a Fourier transform of a uniquely chosen nondecreasing function $G(\lambda)$,

$$C(k) = \int_{-1/2}^{1/2} \cos 2\pi k\lambda \, dG(\lambda) \tag{5.1}$$

where $0 \leqq \lambda \leqq 1/2$. In this context, one refers to λ as a frequency since the function $\cos 2\pi\lambda x$ has period $1/\lambda$ and frequency $1/1/\lambda = \lambda$.

For continuous parameter stochastic processes which are covariance stationary, the covariance function C(h), regarded as a function of h, is also positive semi-definite and can be uniquely represented as a Fourier transform

$$C(h) = \int_0^\infty \cos 2\pi h\lambda \, dG(\lambda) \tag{5.2}$$

where $G(\lambda)$ is a uniquely chosen monotone nondecreasing bounded function and where $0 \leqq \lambda < \infty$. In either the discrete or continuous case the function $G(\lambda)$ is called the spectral function. Expressions corresponding to (5.1) and (5.2) are available for complex valued covariance stationary processes, but our discussion is here limited to real valued processes only. These important results are consequences of Bochner's theorem.

Moreover, in the discrete parameter case, when $\sum_{k=-\infty}^{\infty} |C(k)|$ converges there is a continuous spectral density function, sometimes called the power spectrum,

$$g(\lambda) = 2C(0) + 4 \sum_{k=1}^{\infty} C(k)\cos 2\pi k\lambda. \tag{5.3}$$

This function has as its domain the frequencies which are associated with the corresponding autocovariance function, i.e., $0 \leqq \lambda \leqq 1/2$.

This sometimes can be emphasized by saying that the function $g(\lambda)$ is defined over the frequency domain.

In the continuous parameter case, when $\int_{-\infty}^{\infty} |C(h)|\, dh$ converges, there is a continuous spectral density function or power spectrum

$$g(\lambda) = 2C(0) + 4 \int_{0}^{\infty} C(h)\cos 2\pi\lambda h\, dh. \tag{5.4}$$

The results (5.3) and (5.4) are special cases of the more general Fourier inversion theorem and they permit one to obtain the spectral density function from the autocovariance function.

Since the function $\cos 2\pi\lambda h$ is symmetric about $\lambda = 0$, while $C(k) = C(-k)$ for integers k and $C(h) = C(-h)$ for real numbers h, some writers define the power spectrum as

$$p(\lambda) = C(0) + 2 \sum_{k=1}^{\infty} C(k)\cos 2\pi k\lambda \qquad -1/2 \leq \lambda \leq 1/2$$

for discrete parameter processes or

$$p(\lambda) = C(0) + 2 \int_{0}^{\infty} C(h)\cos 2\pi\lambda h\, dh \qquad -\infty < \lambda < \infty$$

for continuous parameter processes instead of (5.3) and (5.4), respectively. The range of frequencies has been extended so as to include negative integers in the discrete parameter case and negative real numbers in the continuous case, the "power" assigned to the

frequency being one-half that in (5.3) and (5.4), i.e., $p(\lambda) = 1/2\ g(|\lambda|)$.

The preceding discussion provides a dual characterization for the probabilistic properties of covariance stationary processes as they can be examined either in the time domain using the covariance or autocorrelation functions or in the frequency domain through the analysis of the power spectrum. This duality between the time and frequency domains is extremely important for time series analysis and forecasting.

6. MOVING AVERAGE PROCESSES AND SOME EXAMPLES

Consider a sequence $\{\varepsilon(t)$ of mutually uncorrelated random variables having mean 0 and unit variance and where $t = 0, \pm 1, \pm 2, \cdots$ (such a sequence is called white noise); then

$$Z(t) = b_0\varepsilon(t) + b_1\varepsilon(t-1) + \cdots + b_q\varepsilon(t-q), \tag{6.1}$$

where $b_0 \neq 0$ and $b_1, \cdots, b_q$ are arbitrary (real) constants, is called a discrete parameter moving average process of finite order q and is denoted by MA(q). For our purposes, when $b_0, b_1, \cdots$ is an infinite sequence for which $\sum_{j=0}^{\infty} b_j^2 < \infty$ we define the process

$$Z(t) = b_0\varepsilon(t) + b_1\varepsilon(t-1) + \cdots \tag{6.2}$$

to be a discrete parameter moving average process of infinite order.

The mean function of a moving average process of order q is

$$M(t) = E[Z(t)] = E[b_0\varepsilon(t) + b_1\varepsilon(t-1) + \cdots + b_q\varepsilon(t-q)] \tag{6.3}$$

$$= b_0E[\varepsilon(t)] + b_1E[\varepsilon(t-1)] + \cdots + b_qE[\varepsilon(t-q)] = 0$$

since $E[\varepsilon(t)] = 0$ for every t. The variance is

$$C(t) = Var[b_0\varepsilon(t) + b_1\varepsilon(t-1) + \cdots + b_q\varepsilon(t-q)] \tag{6.4}$$

$$= b_0^2Var[\varepsilon(t)] + b_1^2Var[\varepsilon(t-1)] + \cdots + b_q^2Var[\varepsilon(t-q)]$$

$$= \sum_{j=0}^{q} b_j^2,$$

and the autocovariance is

$$C(t_1, t_2) = E[Z(t_1)Z(t_2)] = E[(\sum_{i=0}^{q} b_i\varepsilon(t_1-i))(\sum_{j=0}^{q} b_j\varepsilon(t_2-j))] \tag{6.5}$$

$$= E[\sum_i \sum_j b_ib_j[\varepsilon(t_1-i)\varepsilon(t_2-j)]$$

$$= \sum_i \sum_j b_ib_jE[(\varepsilon(t_1-i)\varepsilon(t_2-j)].$$

When $t_1 = t$ and $t_2 = t+k$, recalling that the $\varepsilon(t)$ have unit variance and are uncorrelated, we find that

$$C(t, t+k) = \sum_i \sum_j b_ib_jE[\varepsilon(t-i)\varepsilon(t+k-j)] \tag{6.6}$$

$$= \sum_{i=0}^{q-k} b_ib_{i+k} \qquad 0 \leq k \leq q$$

$$= 0 \qquad k > q.$$

Since $C(k) = C(t, t+k)$ depends only on k, the MA(q) process is covariance stationary.

The autocorrelation function for the MA(q) process is

$$(6.7) \qquad \rho(k) = \frac{C(k)}{V(0)} = \frac{\sum_{i=0}^{q-k} b_i b_{i+k}}{\sum_{i=0}^{q} b_i^2} \qquad 0 \leq k \leq q$$

$$= 0 \qquad k > q$$

and the power spectrum is

$$(6.8) \quad g(\lambda) = 2\left[\sum_{i=0}^{q} b_i^2\right] + 4 \sum_{k=1}^{q} \left[\sum_{i=0}^{q-k} b_i b_{i+k}\right] \cos 2\pi k \lambda \qquad 0 \leq \lambda \leq 1/2 .$$

Clearly, when we consider an infinite order discrete moving average process (6.2) for which $\sum_{j=0}^{\infty} b_j^2 < \infty$, then Z(t) will be a covariance stationary process with mean $M(t) = 0$, variance $C(0) = V(0) = \sum_{j=0}^{\infty} b_j^2$, and autocovariance $C(k) = \sum_{i=0}^{\infty} b_i b_{i+k}$.

These concepts can be illustrated by using the MA(3) process

$$(6.9) \qquad Z(t) = \varepsilon(t) + .5\,\varepsilon(t-1) + .25\,\varepsilon(t-2) + .125\,\varepsilon(t-3).$$

The mean is 0, of course, and $C(0) = \sum_{i=1}^{3} b_i^2 = 1.328125$, $C(1) = b_0 b_1 + b_1 b_2 + b_2 b_3 = .65625$, $C(2) = .3125$, $C(3) = .125$, while $C(k) = 0$ for $|k| > 3$. For this process $\rho(0) = 1$, $\rho(1) = .50$, $\rho(2) = .23$,

and $\rho(3) = .09$ while $\rho(k) = 0$ for $|k| > 3$. The autocovariance and autocorrelation functions are shown in Figures 1 and 2. Values for negative k are plotted since these are symmetric functions.

The power spectrum for this process is

$$g(\lambda) = 2C(0) + 4 \sum_{k=1}^{\infty} C(k)\cos 2\pi k\lambda$$

$$= 2.65625 + 4[(.65625)\cos 2\pi\lambda$$

$$+ (.3125)\cos 4\pi\lambda + (.125)\cos 6\pi\lambda],$$

$$= 2.65625 + 2.625\cos 2\pi\lambda + 1.25\cos 4\pi\lambda + .50\cos 6\pi\lambda.$$

and is shown in Figure 3.

Returning to the moving average process of finite order q, when the finite dimensional distribution functions of the disturbance terms $\varepsilon(t)$ are multivariate normal, $Z(t)$, as a linear combination of jointly distributed normal random variables, must itself be normally distributed (the sequence of random variables $\varepsilon(t)$, $\varepsilon(t-1)$, $\cdots$, $\varepsilon(t-q)$ in this case is sometimes referred to as Gaussian white noise). Therefore, the finite dimensional distribution functions for the covariance stationary process $Z(t)$ are given by

$$(6.9) \qquad F(z_1, \cdots, z_n; t_1, \cdots, t_n) =$$

$$\int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_n} \frac{1}{(2\pi)^{n/2} |\Sigma(t_1, \cdots, t_n)|^{-1/2}}$$

$$e^{-1/2 z \Sigma^{-1}(t_1, \cdots, t_n) z^T} dz_1 \cdots dz_n$$

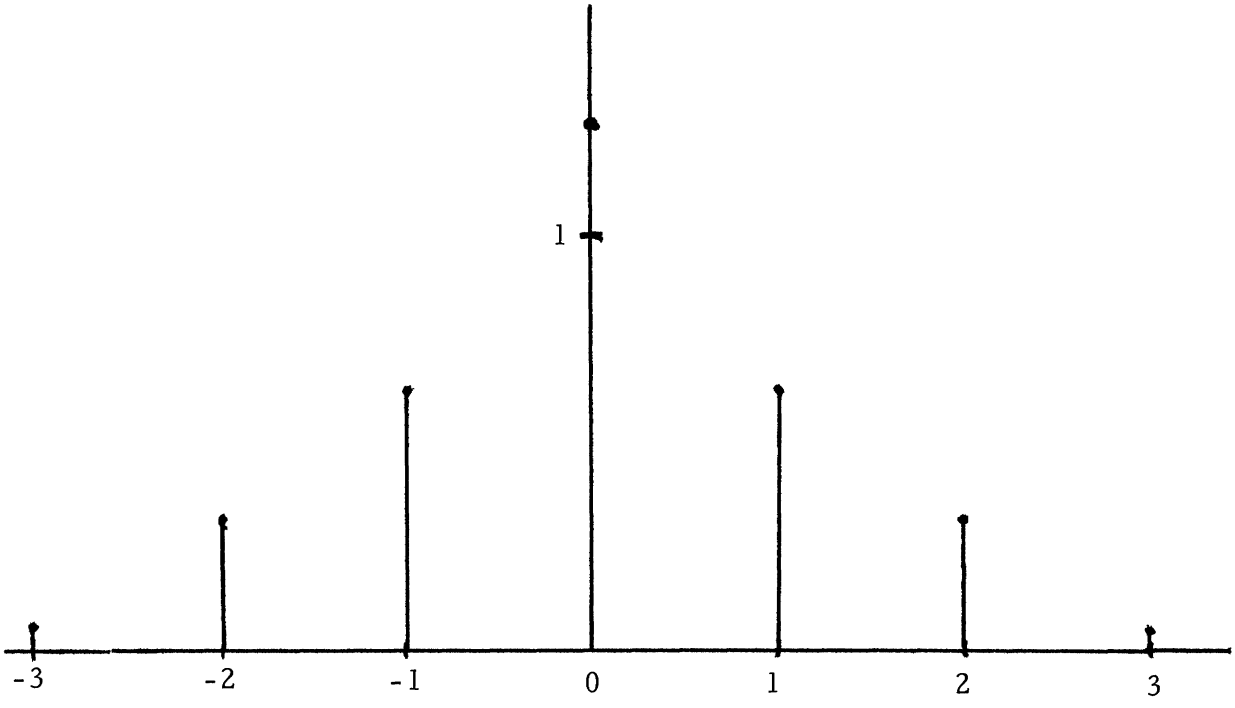

Fig. 1. Autocovariance Function for the MA(3) Process (6.9).

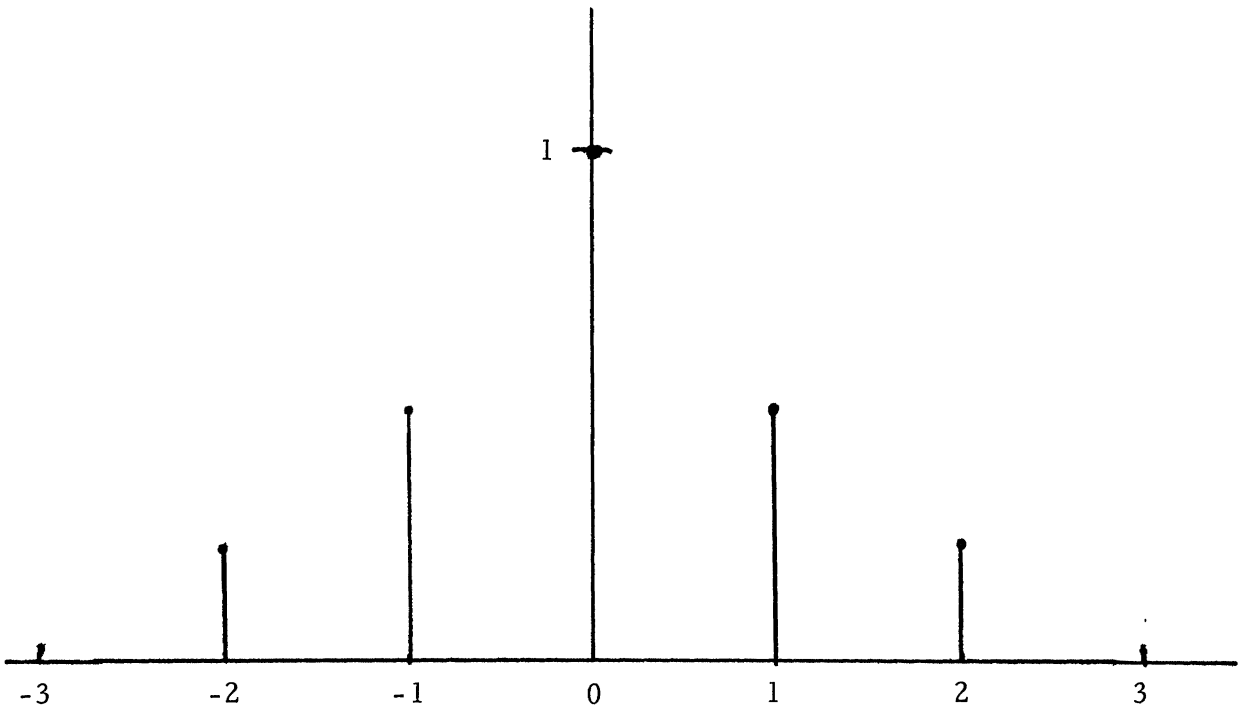

Fig. 2. Autocorrelation Function for the MA(3) Process (6.9).

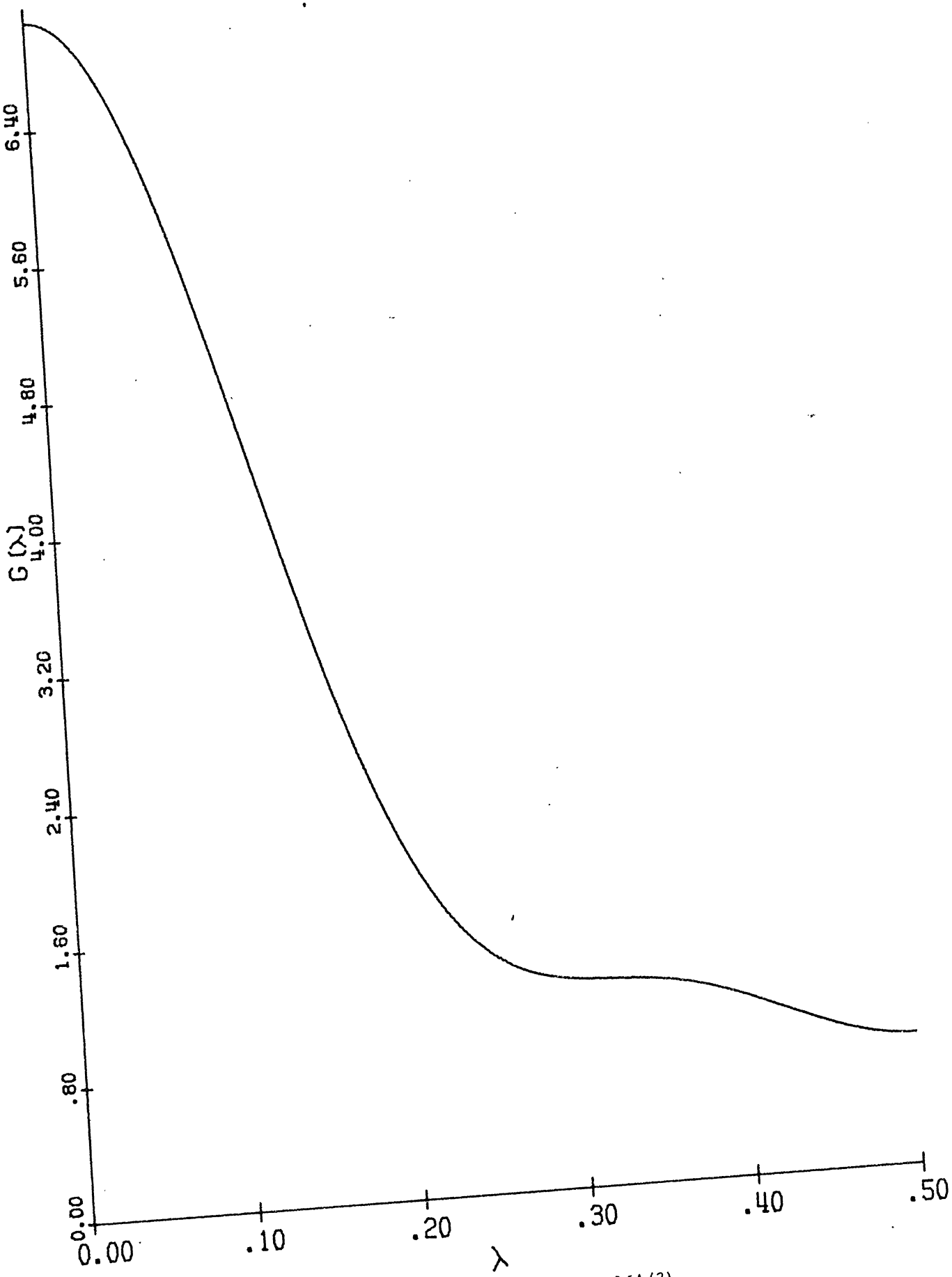

Fig. 3. Power Spectrum for the MA(3) Process (6.9).

where z is the 1 by n vector

$$z = [z_1, \cdots, z_n]$$

and $\Sigma(t_1, \cdots, t_n)$ is the n by n variance-covariance matrix of the random variables $Z(t_1), \cdots, Z(t_n)$,

$$\Sigma(t_1, \cdots, t_n) = \begin{bmatrix} C(t_1, t_1) & \cdots & C(t_1, t_n) \\ \vdots & \vdots & \vdots \\ C(t_n, t_1) & \cdots & C(t_n, t_n) \end{bmatrix}$$

We have seen that for an MA(q) process $C(t_i, t_j) = C(t_i+k, t_j+k)$, so the variance-covariance matrix depends only on the displacement k, i.e.,

$$\Sigma(t_1, \cdots, t_n) = \Sigma(t_1+k, \cdots, t_n + k);$$

it follows that when the random variables $\varepsilon(t)$ are multivariate normal,

$$F(z_1, \cdots, z_n; t_1, \cdots, t_n) = F(z_1, \cdots, z_n; t_1+k, \cdots, t_n+k)$$

and the MA(q) process is strictly stationary as well.

More generally, it can be said that when the families of finite dimensional distributions of a stochastic process are multivariate normal, covariance stationarity of the process implies strict stationarity.

7. AUTOREGRESSIVE PROCESS

Frequently in economic and other applications, it is natural to regard the current value of the time series as being influenced by the lagged values of the process together with a random disturbance term $\varepsilon(t)$ which does not depend on the lagged values. Such a process can be written

$$a_0 Z(t) + a_1 Z(t-1) + \cdots + a_p Z(t-p) = \varepsilon(t) \tag{7.1}$$

where $a_0 \neq 0$, $a_1, \cdots, a_p$ are real constants and $\varepsilon(t)$ for $t = 0, \pm 1, \pm 2, \cdots$ are uncorrelated random variables having mean 0 and unit variance (i.e., white noise). The process (7.1) is called a discrete parameter autoregressive process of order p and is denoted by AR(p). From (7.1) we easily obtain

$$a_0 Z(t) = c_1 Z(t-1) + c_2 Z(t-2) + \cdots + c_p Z(t-p) + \varepsilon(t) \tag{7.2}$$

where $c_j = -a_j$, $j = 1, \cdots, p$, which displays $Z(t)$ as a weighted sum of the lagged $Z(t-j)$ and the disturbance for period t. The equation (7.2) indicates why the process (7.1) is called autoregressive -- the lagged terms in (7.2) play a role that is analogous to explanatory variables in standard single equation regression analysis.

One cannot easily determine the mean, variance, and covariance functions of the process (7.1) as was the case for the MA(q)

process because we do not know the joint distribution of the random variables $Z(t-j)$; only some properties of the disturbances $\varepsilon(t)$ are known in (7.1). Thus it is natural to ask if $Z(t)$ can be expressed only in terms of the random disturbances $\varepsilon(t)$, $\varepsilon(t-1)$, $\cdots$, and under what conditions this is possible. To this end consider (7.1) for $t-1$, $t-2$, $\cdots$, $t-\nu$, $\nu \leq p$,

(7.1)

$$\begin{array}{l}
a_0 Z(t-1) + a_1 Z(t-2) + \cdots + a_p Z(t-p-1) = \varepsilon(t-1) \\
a_0 Z(t-2) + a_1 Z(t-3) + \cdots + a_p Z(t-p-2) = \varepsilon(t-2) \\
\quad\vdots \\
a_0 Z(t-\nu) + a_1 Z(t-\nu-1) + \cdots + a_p Z(t-p-\nu) = \varepsilon(t-\nu)
\end{array}$$

Now upon multiplying (7.1) by $b_0 \neq 0$ and each of the above equations by b_1, b_2, $\cdots$, b_ν, respectively, we obtain the system of equations

$$\begin{array}{l}
a_0 b_0 Z(t) + a_1 b_0 Z(t-1) + \cdots + a_p b_0 Z(t-p) = b_0 \varepsilon(t) \\
a_0 b_1 Z(t-1) + a_1 b_1 Z(t-2) + \cdots + a_p b_1 Z(t-p-1) = b_1 \varepsilon(t-1) \\
\quad\vdots \\
a_0 b_\nu Z(t-\nu) + a_1 b_\nu Z(t-\nu-1) + \cdots + a_p b_\nu Z(t-p-\nu) = b_\nu \varepsilon(t-\nu)
\end{array}$$

Adding these equations together and transposing to the right hand side those $Z(t)$ beyond ν periods in the past gives

$$\sum_{j=0}^{\nu} d_j Z(t-j) = \sum_{j=0}^{\nu} b_j \varepsilon(t-j) + R_Z(\nu, p) \tag{7.3}$$

where

$$d_j = \sum_{k=0}^{j} a_k b_{j-k} \qquad j = 0, 1, 2, \cdots, \nu$$

$$d_j = \sum_{k=0}^{\nu} a_{j+k-\nu} b_{\nu-k} \qquad j = \nu+1,\ \nu+2,\ \cdots,\ \nu + (p-\nu)$$

$$d_j = \sum_{k=j}^{p+\nu} a_{k-\nu} b_{\nu+j-k} \qquad j = p+1,\ p+2,\ \cdots,\ p+\nu$$

and where the remainder term is

$$R_Z(\nu, p) = - \sum_{j=\nu+1}^{p+\nu} d_j Z(t-j).$$

For example, for $j = 1, \cdots, \nu$ we get

$$\begin{aligned} d_0 &= a_0 b_0 \\ d_1 &= a_0 b_1 + a_1 b_0 \\ d_2 &= a_0 b_2 + a_1 b_1 + a_2 b_0 \\ &\cdots \\ d_\nu &= a_1 b_\nu + a_1 b_{\nu-1} + \cdots + a_\nu b_0 \end{aligned} \tag{7.4}$$

Clearly, $Z(t-1)$, $Z(t-2)$, $\cdots$, $Z(t-\nu)$ will not appear in (7.3) when

$$d_j = \sum_{k=1}^{j} a_k b_{j-k} = 0, \qquad j = 1, 2, \cdots, \nu. \tag{7.5}$$

When this is the case we have

$$d_0 Z(t) = \sum_{j=0}^{\nu} b_j \varepsilon(t-j) + R_Z(\nu, p) \tag{7.6}$$

which expresses Z(t) as a weighted sum of the random disturbances for the immediately preceding ν periods and a weighted sum of the Z(t) for past periods beyond ν. Now as ν increases, (7.6) indicates that Z(t) is influenced by successively more disturbance terms, together with a weighted sum of previous values of the time series for successively more remote periods of time. Consequently, in the limit we obtain

$$d_0 Z(t) = \sum_{j=0}^{\infty} b_j \varepsilon(t-j) + R_Z(\infty)$$

where

$$R_Z(\infty) = \lim_{\substack{r \to \infty \\ \nu \leq p}} R_Z(\nu, p).$$

For purposes of analyzing economic time series the "infinitely remote past history" reflected in $R_Z(\infty)$ can be assumed to be unimportant. If these influences are ignored we arrive at the representation

$$a_0 Z(t) = \sum_{j=0}^{\infty} b_j \varepsilon(t-j) \tag{7.7}$$

or, when $a_0 = 1$,

$$Z(t) = \sum_{j=0}^{\infty} b_j \varepsilon(t-j) \tag{7.8}$$

and we have Z(t) expressed as a moving average process of infinite order.

Under the condition that $\sum_{j=0}^{\infty} b_j^2 < \infty$, we know from the previous section that the mean function of the autoregressive process (7.1) is 0, its variance is $C(0) = \sum_{j=0}^{\infty} b_j^2/a_0^2$, and its autocovariance function is $C(k) = \sum_{i=0}^{\infty} b_i b_{i+k}/a_0^2$. The power spectrum of the process is

$$(7.9) \qquad g(\lambda) = 2\Big[\sum_{i=0}^{\infty} b_i^2/a_0^2\Big] + 4 \sum_{k=1}^{\infty} \Big[\sum_{i=1}^{\infty} b_i b_{i+k}/a_0^2\Big] \cos 2\pi k\lambda .$$

To complete this discussion we should consider how the b_j in (7.8) can be obtained from the a_j in (7.1). One way to proceed is to use the equations in (7.5). For example, if $j = 1$ we obtain b_1 from the condition $a_0 b_1 + a_1 b_0 = 0$ as

$$b_1 = -\frac{a_1 b_0}{a_0} .$$

For $j = 2$, we obtain b_2 from the condition $a_0 b_2 + a_1 b_1 + a_2 b_0 = 0$ as

$$b_2 = -\frac{a_1 b_1 - a_2 b_0}{a_0} .$$

This direct recursive procedure becomes extremely involved, however, and infinitely many equations must be solved.

An alternate method for determining the b_j is to use the generating functions:

$$A(x) = \sum_{k=0}^{\infty} a_k x^k \qquad B(x) = \sum_{k=0}^{\infty} b_k x^k .$$

Now

$$(7.10) \qquad A(x)B(x) = \Big(\sum_{k=0}^{\infty} a_k x^k\Big)\Big(\sum_{k=0}^{\infty} b_k x^k\Big) = \sum_{j=0}^{\infty} d_j x^j$$

where d_j are given in (7.3). If $d_j = 0$ for $j \geq 1$, we have

$$(7.11) \qquad A(x)B(x) = d_0 = a_0 b_0 .$$

Conversely, if (7.11) holds for all x, this requires that $d_j = 0$ for $j \geq 1$. Thus we can say that the equations (7.5) are satisfied if and only if the product of the generating functions gives $A(x)B(x) = a_0 b_0$. If this is the case, we get

$$B(x) = \frac{a_0 b_0}{A(x)} ,$$

so that B(x) is a ratio of two polynomials. As a rational function, the coefficients b_j of B(x) can be obtained from the partial fraction expansion of B(x). For example, suppose that A(x) has p distinct roots $x_1, \ldots, x_p$; then the partial fraction expansion of B(x) is

$$(7.12) \qquad B(x) = \sum_{k=1}^{p} A_k (x_k - x)^{-1} .$$

Recalling that $B(x) = \sum_{k=0}^{\infty} b_k x^k$, we can differentiate B(x) successively and evaluate these derivatives at x = 0, getting

$$B^{(j)}(0) = j!b_j \qquad j = 0, 1, 2, \cdots, \tag{7.13}$$

But since B(x) has the representation (7.12), successive differentiation of the right-hand side of (7.12) gives

$$\frac{d^j}{dx^j}\Big[\sum_{k=1}^{p} A_k(x_k - x)^{-1}\Big] = \sum_{k=1}^{p} j!A_k(x_k - x)^{-(j+1)} .$$

Therefore, upon evaluating these derivatives at x = 0, since

$$B^{(j)}(0) = \frac{d^j}{dx^j}\Big[\sum_{k=1}^{p} A_k(x_k - x)^{-1}\Big]_{x=0} =$$

we obtain

$$j!b_j = \sum_{k=1}^{p} j!A_k x_k^{-(j+1)}$$

or

$$b_j = \frac{A_1}{x_1^{j+1}} + \frac{A_2}{x_2^{j+1}} + \cdots + \frac{A_p}{x_p^{j+1}}, \qquad j = 0, 1, 2, \cdots, \tag{7.14}$$

Thus to determine the b_j in (7.4), instead of solving infinitely many equations recursively, we need only obtain the roots of the polynomial A(x) and use (7.14) to determine the required coefficients b_j in the infinite order moving average representation of the underlying auto-regressive process. Moreover, even when the roots of the polynomial A(x) are not distinct a similar representation for the coefficients b_j can be obtained.

More importantly, the above discussion gives an insight into the stationarity of the autoregressive process. We have seen that an AR process is stationary provided that $\sum_{k=1}^{\infty} b_k^2 < \infty$. Suppose one of the roots x satisfies $|x_i| > 1$; then from (7.14) it can be seen that $|b_j| \to \infty$ since $\lim |x_i|^{j+1} = 0$. In other words, the AR(p) process is covariance stationary if and only if the roots of the polynomial A(x) all lie outside the unit circle.

The autocorrelations associated with a covariance stationary process satisfy an important set of relationships called the Yule-Walker equations. These are obtained by multiplying (7.1) by Z(t-k) for $k > 0$,

$$a_0 Z(t)Z(t-k) + \cdots + a_p Z(t-p)Z(t-k) = \varepsilon(t)Z(t-k). \tag{7.15}$$

Upon taking expected values we get

$$a_0 E[Z(t)Z(t-k)] + \cdots + a_p E[Z(t-p)Z(t-k)] = E[\varepsilon(t)Z(t-k)]. \tag{7.16}$$

Since $k > 0$, Z(t-k) refers to a past period whereas $\varepsilon(t)$ is the current disturbance, so $\varepsilon(t)$ and Z(t-k) in this case are uncorrelated and we have $E[\varepsilon(t)Z(t-k)] = 0$. Also, since $C(k) = E[Z(t)Z(t-k)]$, we see that (7.16) becomes

$$a_0 C(k) + a_1 C(k-1) + \cdots + a_p C(k-p) = 0, \qquad k > 0. \tag{7.17}$$

This system of difference equations involving the C(j) can be used to

characterize the autocorrelation function of a covariance stationary AR(p) process since if we divide (7. 17) by V(0) we obtain

$$a_0\,\rho(k) + a_1\,\rho(k-1) + \cdots + a_p\,\rho(k-p) = 0. \tag{7.18}$$

The latter system consists of the Yule-Walker equations and they play an important role in estimation using data from given time series.

Moreover, when the autocorrelation function is given, we can regard the system (7. 18) as determining the coefficients a_j in the autoregressive scheme (7. 1). This system can be written in matrix form as

$$\begin{bmatrix} 1 & \rho(1) & \cdots & \rho(1-p) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \rho(p-1) & & & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_p \end{bmatrix} = -a_0 \begin{bmatrix} \rho(1) \\ \cdot \\ \cdot \\ \cdot \\ \rho(p) \end{bmatrix} . \tag{7.19}$$

For example, for the AR(1) process where $a_0 = -1$ the system (7. 19) becomes

$$[\,1\,]\,[a_1(1)\,] = [\rho(1)\,]$$

or

$$a_1(1) = \rho(1),$$

where $a_1(1)$ is intended to indicate the order of the corresponding autoregressive process. For an AR(2) process where $a_0 = -1$ the system

becomes

$$\begin{bmatrix} 1 & \rho(1) \\ \rho(1) & 1 \end{bmatrix} \begin{bmatrix} a_1(2) \\ a_2(2) \end{bmatrix} = \begin{bmatrix} \rho(1) \\ \rho(2) \end{bmatrix};$$

we get

$$\begin{bmatrix} a_1(2) \\ a_2(2) \end{bmatrix} = \begin{bmatrix} 1 & \rho(1) \\ \rho(1) & 1 \end{bmatrix}^{-1} \begin{bmatrix} \rho(1) \\ \rho(2) \end{bmatrix},$$

and finally

$$a_1(2) = \frac{\rho(1) - \rho(1)\rho(2)}{1 - [\rho(1)]^2}$$

$$a_2(2) = \frac{\rho(2) - [\rho(1)]^2}{1 - [\rho(1)]^2} .$$

Continuing in this fashion the coefficients a_j of the AR(p) process (7.1) can be considered as functions $a_j(p)$ of the order of the autoregressive scheme.

In this interpretation the $a_1(1)$, $a_2(2)$, $\cdots$, $a_p(p)$ are called partial autocorrelations. We will refer to these concepts again in a later section.

8. AUTOREGRESSIVE MOVING AVERAGE PROCESSES

In the AR(p) process (7.1) the random disturbances $\varepsilon(t)$ are assumed to be uncorrelated random variables with mean 0 and unit

variance. The probability structure of these disturbances is so simple that it is natural, especially in economic time series where the disturbances are not uncorrelated from period to period, to consider more general properties for the disturbance or error terms. Accordingly, we replace the error terms $\varepsilon(t)$ in (7. 1) by a moving average of order q and obtain the autoregressive moving average process, denoted ARMA(p, q); namely,

$$(8.1) \qquad a_0 Z(t) + a_1 Z(t-1) + \cdots + a_p Z(t-p) = b_0 \varepsilon(t) + b_1 \varepsilon(t-1) + \cdots + b_q \varepsilon(t-q).$$

Two questions immediately arise: are these processes stationary? What are their means, variances, autocovariances, and autocorrelations? To consider the first question, rewrite (8. 1) in the form

$$(8.2) \qquad a_0 Z(t) + a_1 Z(t-1) + \cdots + a_p Z(t-p) = \varepsilon^*(t)$$

where

$$(8.3) \qquad \varepsilon^*(t) = b_0 \varepsilon(t) + \cdots + b_q \varepsilon(t-q).$$

Now whenever the right hand side of (8. 2) is a covariance stationary process (as is the case for a MA(q) process) the arguments given in the preceding section establishing the covariance stationarity of the AR(p) process will also establish the covariance stationarity of (8. 1).

In other words, under the assumption that the disturbances or errors are covariance stationary, (8.1) will be covariance stationary provided the roots of the polynomial A(x) referred to in the previous section lie outside the unit circle.

Returning to a consideration of the second question posed above, if we use the ARMA(p, q) process (8.2) and the arguments of the previous section we see that we can express Z(t) in the form

$$a_0 Z(t) = \sum_{j=0}^{\infty} b_j^* \varepsilon^*(t-j), \tag{8.4}$$

where $\varepsilon^*(t)$ is given in (8.3). From the representation (8.4) we see that the mean function of the ARMA(p, q) process is M(t) = 0. The variance, autocovariance, and autocorrelation functions can be obtained explicitly but each would be more lengthy than those we dealt with earlier because one must account for the fact that the $\varepsilon^*(t)$ are no longer uncorrelated random variables.

An alternate characterization for these functions can be developed implicitly using a system of difference equations that they must satisfy. To obtain this system, multiply (8.1) by Z(t-k) for $k \geq 0$, getting

$$\begin{aligned} a_0 Z(t)Z(t-k) + a_1 Z(t-1)Z(t-k) + \cdots + a_p Z(t-p)Z(t-k) = \\ b_0 \varepsilon(t)Z(t-k) + b_1 \varepsilon(t-1)Z(t-k) + \cdots + b_q \varepsilon(t-q)Z(t-k). \end{aligned} \tag{8.5}$$

Upon taking expectations we obtain

$$(8.6)\quad a_0E[Z(t)Z(t-k)] + a_1E[Z(t-1)Z(t-k)] + \cdots + a_pE[Z(t-p)Z(t-k)] =$$
$$b_0E[\varepsilon(t)Z(t-k)] + b_1E[\varepsilon(t-1)Z(t-k)] + \cdots + b_qE[\varepsilon(t-q)Z(t-k)].$$

Since $Z(t)$ is correlated with $\varepsilon(t)$, $\varepsilon(t-1)$, $\cdots$, $\varepsilon(t-q)$, the cross-product expectations involving $Z(t)$ and $\varepsilon(t)$ beyond the current period are not necessarily zero as is the case for the AR(p) model. These expectations are called the cross-covariance which we denote

$$C_{\varepsilon Z}(j-i) = E[\varepsilon(t-i)Z(t-j)],$$

and, when $j = 0$, we get

$$C_{\varepsilon Z}(-i) = E[\varepsilon(t-i)Z(t)].$$

Note that when $i < j$ we have $C_{\varepsilon Z}(j-i) = 0$ because we are then dealing with $Z(t)$ for values of t previous to the time of the given disturbance and in this case the $Z(t)$ and $\varepsilon(t)$ are uncorrelated. Also, when $i \geq j$, $C(j-i)$ will not be zero.

Using this notation, (8.6) can be written for $k \geq 0$ as

$$(8.7)\quad a_0C(k) + a_1C(k-1) + \cdots + a_pC(k-p) =$$
$$b_0C_{\varepsilon Z}(k) + b_1C_{\varepsilon Z}(k-1) + \cdots + b_qC_{\varepsilon Z}(k-q).$$

For example, taking $k = 0, 1, 2, \cdots, q$ successively in (8.7) we obtain

$$a_0C(0)+a_1C(1)+\cdots+a_pC(p) = b_0C_{\varepsilon Z}(0)+b_1C_{\varepsilon Z}(-1)+\cdots$$

$$+b_qC_{\varepsilon Z}(-q)$$

$$a_0C(1)+a_1C(2)+\cdots+a_pC(1-p) = b_1C_{\varepsilon Z}(0)+b_2C_{\varepsilon Z}(-1)+\cdots$$

$$+b_qC_{\varepsilon Z}(1-q)$$

(8.8) $$a_0C(2)+a_1C(3)+\cdots+a_pC(2-p) = b_2C_{\varepsilon Z}(0)+b_3C_{\varepsilon Z}(-1)+\cdots$$

$$+b_qC_{\varepsilon Z}(2-q)$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$a_0C(q)+a_1C(q-1)+\cdots+a_pC(q-p) = b_qC_{\varepsilon Z}(0).$$

Whenever $k > q$, we get

(8.9) $$a_0C(k) + a_1C(k+1) + \cdots + a_pC(k-p) = 0$$

since in this case all cross-covariances appearing on the right-hand side of (8.7) are zero.

The system of difference equations represented in (8.8) and (8.9) implicitly defines the autocovariance functions. The autocorrelation functions $\rho(k)$ satisfy the same system of equations and are also implicitly defined by means of them.

9. SIMPLE EXPONENTIAL SMOOTHING FORECASTING MODELS

Suppose a components model for a time series is given by

$$Y(t) = TC(t) + e(t) \tag{9.1}$$

where

$$TC(t) = s_0 + s_1 t + \cdots + s_q t^q = P(t) \tag{9.2}$$

and where s_j is real, $s_q \neq 0$, and e(t) is an irregular or random component representing uncorrelated random variables with mean zero. That is, we have an additive model in which the trend-cyclical component is assumed to be a polynomial of order q. One wishes to estimate P(t) in order to develop a forecasting model.

When the errors e(t) are uncorrelated with mean zero and constant variance σ_e^2, given observations on Y(t) for time periods 0, 1, $\cdots$, t (we call the first period for which we have observations time period 0 and denote the current period by t), then the s_j in (9.2) could be estimated by ordinary least squares, the estimates s_j being determined so as to minimize

$$\sum_{j=0}^{t} (Y(t-j) - P(t-j))^2. \tag{9.3}$$

More generally, in order to develop estimates s_j of the coefficients in (9.2), when the variances of the errors or disturbances are not constant from observation to observation, we can minimize the

weighted sum of squares

$$(9.4) \qquad \sum_{j=0}^{t} \frac{1}{\sigma^2(t-j)} (Y(t-j) - P(t-j))^2,$$

where

$$\sigma^2(t-j) = E[e^2(t-j)].$$

These are standard procedures and if in addition the errors are normally distributed, maximum likelihood estimators can be obtained and the conventional large sample inference procedures would be available as well.

For many economic time series the variances of the error terms decrease as one approaches the current time period. If this were not the case and the variances were increasing in time, forecasts for future periods, even those for one period ahead, would be virtually useless in practice because they would be anticipated to differ so widely from the actual future values for the time series.

Returning to the components model (9.1), it is therefore natural to think in terms of discounting previous observations of Y(t) more heavily than the current observation, or equivalently, to think in terms of discounting previous errors more heavily than the current error. A convenient way of doing this is to assign weights which decrease as a geometric sequence,

$$(9.5) \qquad \beta^0, \ \beta^1, \ \beta^2, \ \ldots, \ \beta^t.$$

In this case the components model can be considered as

$$Y(t-j) = P(t-j) + e(t-j) \qquad j = 0, 1, \cdots, t \tag{9.6}$$

where it is assumed that the e(t-j) are uncorrelated with mean zero but have variance proportional to β^{-j} for j = 1, 2, ···, t, say $\sigma^2(t-j) = 1/\sigma^2\beta^j$, where σ^2 is a constant of proportionality. Since the variance is no longer constant we multiply (9.6) by $1/\sigma(t-j)$, obtaining

$$\frac{Y(t-j)}{\sigma(t-j)} = \frac{P(t-j)}{\sigma(t-j)} + \frac{e(t-j)}{\sigma(t-j)}\ , \tag{9.7}$$

which we can write as

$$Y^*(t-j) = P^*(t-j) + \varepsilon(t-j), \tag{9.8}$$

where the $\varepsilon(t-j)$ are uncorrelated with mean zero and, it should be noted, have unit variance, i.e., are white noise. Thus the ordinary least squares criterion (9.3) applied to this representation is

$$\begin{aligned} \sum_{j=0}^{t} \varepsilon^2(t-j) &= \sum_{j=0}^{t} (Y^*(t-j) - P^*(t-j))^2 \\ &= \sum_{j=0}^{t} \left[\frac{Y(t-j)}{\sigma(t-j)} - \frac{P(t-j)}{\sigma(t-j)}\right]^2 \\ &= \sigma^2 \sum_{j=0}^{t} \beta^j (Y(t-j) - P(t-j))^2 . \end{aligned} \tag{9.9}$$

We call the minimization of (9.9) with respect to the coefficient s_j of (9.2) the discounted least squares criterion.

To illustrate the use of this criterion, we consider the case in which $P(t) = s_0$ and estimate s_0 by choosing that value $\hat{s}_0$ which minimizes

$$\sum_{j=0}^{t} (Y(t-j) - s_0)^2 . \tag{9.10}$$

Upon differentiating (9.10) with respect to s_0, setting the derivative equal to zero, and solving for $\hat{s}_0$, we get

$$\hat{s}_0 = \frac{\sum_{j=0}^{t} \beta^j Y(t-j)}{\sum_{j=0}^{t} \beta^j} = \frac{\sum_{j=0}^{t} (1-\beta)\beta^j Y(t-j)}{1-\beta^{t+1}} \tag{9.11}$$

since

$$\sum_{j=0}^{t} \beta^j = \frac{1-\beta^{t+1}}{1-\beta} .$$

The factor $1-\beta$ is frequently denoted by α and called the exponential smoothing constant. In practice the choice of the smoothing constant α is based upon the user's judgment regarding the weight he wishes to assign to the most recent observation, rather than on the application of appropriate methods of statistical inference.

Provided $|\beta| < 1$, (9.11) can be rewritten in the limiting form

$$\hat{s}_0 = \sum_{j=0}^{t-1} \alpha\beta^j Y(t-j) + \beta^t S(0) \tag{9.12}$$

where

$$S(0) = \alpha Y(0) + \alpha\beta Y(-1) + \alpha\beta^2 Y(-2) + \cdots .$$

One sees that when t is large (a large number of time series observations is available), the term S(0) in (9.12) depends only on observations of the time series from the remote past and in such a way that the observations beyond Y(0) are discounted successively more heavily. Moreover, in (9.12), S(0) is itself discounted by β^t, so when t is large one may ignore the last term on the right in (9.12) and use the estimate

$$\hat{s}_0 = \sum_{j=0}^{t-1} \alpha\beta^j Y(t-j). \tag{9.13}$$

In (9.13), when t is allowed to assume integer values successively, we interpret $\hat{s}_0$ as a function of t and write $\hat{s}_0(t)$ rather than s_0. In other words, for $t = 1, 2, 3, \cdots$ we have

$$\hat{s}_0(t) = \sum_{j=0}^{t-1} \alpha\beta^j Y(t-j) \tag{9.14}$$

$$= \alpha Y(t) + \sum_{j=1}^{t-1} \alpha\beta^j Y(t-j)$$

$$= \alpha Y(t) + \beta \sum_{j=1}^{t-1} \alpha\beta^{j-1} Y(t-j).$$

Changing the index of summation appearing in the last term above to $i = j-1$ gives

$$\hat{s}_0(t) = \alpha Y(t) + \beta \sum_{i=0}^{t-2} \alpha\beta^i Y(t-1-j).$$

From (9.14), however, when t is replaced by t-1, we obtain

$$\hat{s}_0(t-1) = \sum_{i=0}^{t-2} \alpha\beta^i Y(t-1-j), \tag{9.15}$$

and substituting (9.15) into (9.14) leads to

$$\hat{s}_0(t) = \alpha Y(t) + \beta \hat{s}_0(t-1) . \tag{9.16}$$

Rather than denoting the estimate (9.16) of the constant term in (9.1) by $\hat{s}_0(t)$, the alternate notation $S^{(1)}(t)$ is frequently used and expressing (9.14) in this notation gives

$$S^{(1)}(t) = \alpha Y(t) + \beta S^{(1)}(t-1) \qquad \beta = 1 - \alpha \quad 0 < \alpha \leq 1. \tag{9.17}$$

The equation gives the <u>first order exponentially smoothed value of Y(t);</u> it provides a simple recursive procedure for updating the smoothed values $S^{(1)}(t)$ and has an appealing intuitive justification. If we "unfold" (9.17) by substituting $S^{(1)}(t)$, $S^{(1)}(t-1)$, $\cdots$, as expressed by (9.17) successively in this equation, we see that $S^{(1)}(t)$ can be written as a weighted sum of the observations Y(t), Y(t-1), $\cdots$, with the weights

decreasing in a geometric progression as the observations refer to successively more distant time periods.

Finally, for forecasting purposes, since the time series is represented as

$$Y(t) = s_0 + e(t),$$

the mean function of Y(t) under the assumed properties for the disturbance term e(t) is

(9.18) $$M(t) = E[Y(t)] = s_0.$$

Thus an estimate $\hat{Y}(t+\tau)$ of the expected value of Y(t) for τ periods ahead is given by

(9.19) $$\hat{Y}(t+\tau) = \hat{M}(t+\tau) = \hat{s}_0(t) = S^{(1)}(t) .$$

Note that in this case, since the trend-cyclical component is given by a polynomial of degree zero, $\hat{Y}(t+\tau)$ does not depend on τ. In other words, we have the same forecast for any period $t+\tau$ in the future and this forecast is changed only when the next observation on the time series becomes available.

The results are extended in the next section to time series in which the trend-cyclical component is not restricted to a polynomial of degree zero but can be of any degree q.

10. HIGHER ORDER EXPONENTIAL SMOOTHING FORECASTING MODELS

Consider the more general case in which the trend-cyclical component is a polynomial of degree q, and for $j = 0, 1, 2, \cdots, t$, express (9.2) in the form

$$P(t-j) = \sum_{i=0}^{q} s_i(t)(-j)^i/i! \tag{10.1}$$

Estimating the coefficients

$$s_0(t),\ s_1(t),\ \cdots,\ s_q(t) \tag{10.2}$$

in (10.1) by minimizing (9.9) gives estimates $\hat{s}_0(t),\ \hat{s}_1(t),\ \cdots,\ \hat{s}_q(t)$ which satisfy the system of equations

$$\begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1,q+1} \\ m_{21} & m_{22} & \cdots & m_{2,q+1} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ m_{q+1,1} & m_{q+1,2} & \cdots & m_{q+1,q+1} \end{bmatrix} \begin{bmatrix} \hat{s}_0(t) \\ \hat{s}_1(t) \\ \cdot \\ \cdot \\ \cdot \\ \hat{s}_q(t) \end{bmatrix} = \begin{bmatrix} S^{(1)}(t) \\ S^{(2)}(t) \\ \cdot \\ \cdot \\ \cdot \\ S^{(q+1)}(t) \end{bmatrix} \tag{10.3}$$

where

$$\begin{aligned} S^{(1)}(t) &= \alpha Y(t) + \beta S^{(1)}(t-1) \\ S^{(2)}(t) &= \alpha S^{(1)}(t) + \beta S^{(2)}(t-1) \\ &\cdot \quad\quad\quad\quad \cdot \quad\quad\quad\quad \cdot \\ &\cdot \quad\quad\quad\quad \cdot \quad\quad\quad\quad \cdot \\ &\cdot \quad\quad\quad\quad \cdot \quad\quad\quad\quad \cdot \\ S^{(q+1)}(t) &= \alpha S^{(q)}(t) + \beta S^{(q+1)}(t-1) \end{aligned} \tag{10.4}$$

and

$$(10.5) \qquad m_{ij+1} = \frac{(-1)^j \alpha^i}{j!(i-1)!} \sum_{k=0}^{\infty} \frac{k^j \beta^k (i-1+k)!}{k!} \qquad \begin{matrix} i \geq 1 \\ j \geq 0 \end{matrix}$$

The $S^{(j)}(t)$ in (10.4) is called the j^{th} order exponentially smoothed value of the time series Y(t). One can also see in (10.4) a simple recursive pattern similar to that in first order exponential smoothing as given in (9.17). Indeed in (10.4) one is smoothing successively time series consisting in turn of other lower order smoothed values of the time series Y(t). As with first order exponential smoothing, initial values $S^{(1)}(0)$, $S^{(2)}(0)$, $\cdots$, $S^{(q+1)}(0)$ must be chosen.

As an example of these concepts consider the case q = 1, in which the trend-cyclical component of the time series Y(t) has a linear trend and can be represented as

$$TC(t) = s_0 + s_1 t = P(t),$$

which is a special case of (9.2), or, alternately, as

$$TC(t+\tau) = s_0(t) + s_1(t)\tau = P(t+\tau),$$

which is a special case of (10.1) where $j = -\tau$. Then corresponding to (10.3), (10.4), and (10.5) we have

$$\begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} \hat{s}_0(t) \\ \hat{s}_1(t) \end{bmatrix} = \begin{bmatrix} S^{(1)}(t) \\ S^{(2)}(t) \end{bmatrix} \tag{10.6}$$

where

$$\begin{aligned} S^{(1)}(t) &= \alpha Y(t) + \beta S^{(1)}(t-1) \\ S^{(2)}(t) &= \alpha S^{(1)}(t) + \beta S^{(2)}(t-1) \end{aligned} \tag{10.7}$$

and the m_{ij+1} in this case are (with, respectively, $i = 1$, $j = 0$; $i = 1$, $j = 1$; $i = 2$, $j = 0$; $i = 2$, $j = 1$):

$$m_{11} = \alpha \sum_{k=0}^{\infty} \frac{\beta^k k!}{k!} = \alpha\left(\frac{1}{\alpha}\right) = 1$$

$$m_{12} = -\alpha \sum_{k=0}^{\infty} k\beta^k = -\alpha\left(\frac{\beta}{\alpha^2}\right) = -\frac{\beta}{\alpha}$$

$$m_{21} = \alpha^2 \sum_{k=0}^{\infty} \frac{\beta^k (1+k)!}{k!} = \alpha^2 \sum_{k=0}^{\infty} (1+k)\beta^k = 1 \tag{10.8}$$

$$m_{22} = -\alpha^2 \sum_{k=0}^{\infty} \frac{k\beta^k (1+k)!}{k!} = -\alpha^2 \sum_{k=0}^{\infty} (1+k)k\beta^k$$

$$= -\alpha^2 \left[\frac{\beta(1+\beta)}{\alpha^3} + \frac{\beta}{\alpha^2}\right] = -\frac{2\beta}{\alpha} .$$

Then (10.6) becomes

$$\begin{bmatrix} 1 & -\frac{\beta}{\alpha} \\ 1 & -\frac{2\beta}{\alpha} \end{bmatrix} \begin{bmatrix} \hat{s}_0(t) \\ \hat{s}_1(t) \end{bmatrix} = \begin{bmatrix} S^{(1)}(t) \\ S^{(2)}(t) \end{bmatrix}.$$

Therefore,

$$\begin{bmatrix} \hat{s}_0(t) \\ \hat{s}_1(t) \end{bmatrix} = \begin{bmatrix} 1 & -\frac{\beta}{\alpha} \\ 1 & -\frac{2\beta}{\alpha} \end{bmatrix}^{-1} \begin{bmatrix} S^{(1)}(t) \\ S^{(2)}(t) \end{bmatrix}$$

$$\begin{bmatrix} \hat{s}_0(t) \\ \hat{s}_1(t) \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ \frac{\alpha}{\beta} & -\frac{\alpha}{\beta} \end{bmatrix} \begin{bmatrix} S^{(1)}(t) \\ S^{(2)}(t) \end{bmatrix}$$

or

$$\hat{s}_0(t) = 2S^{(1)}(t) - S^{(2)}(t) \tag{10.9}$$

$$\hat{s}_1(t) = \frac{\alpha}{\beta}(S^{(1)}(t) - S^{(2)}(t)). \tag{10.10}$$

For a given smoothing constant α, together with given initial values $S^{(1)}(0)$ and $S^{(2)}(0)$, $\hat{s}_0(t)$ and $\hat{s}_1(t)$ are updated whenever a new observation on the time series $Y(t)$ becomes available. The updating process utilizes the recursive expressions (10.7); one first obtains updated values of $S^{(1)}(t)$ and $S^{(2)}(t)$ and these are then used in (10.9) and (10.10) to obtain the revised estimates $\hat{s}_0(t)$ and $\hat{s}_1(t)$.

In the same way as is illustrated for the case $q = 1$, (10.3), (10.4), and (10.5) in principle can be used to determine the estimates $\hat{s}_0(t)$, $\hat{s}_1(t)$, $\cdots$, $\hat{s}_q(t)$ as linear combinations of the various exponentially smoothed values $S^{(1)}(t)$, $S^{(2)}(t)$, $\ldots$, $S^{(q+1)}(t)$. The result that, except for exceptional values of α, the linear equation system (10.3) can be solved is called the fundamental theorem of exponential smoothing (see Brown [4]).

However, for exponential smoothing of order higher than three, serious computational limitations occur if one attempts to use (10.5) to determine the m_{ij+1} because they depend on α, and for each new choice of α, a new calculation must be made. Furthermore, progressively more complex closed form expressions arise for the infinite sums involved in (10.5). Cogger [6] has provided a simplified procedure for determining the elements m_{ij+1}.

Turning now to the problem of developing τ-step ahead forecasts when the trend-cyclical component is a polynomial $P(t)$ of degree q, the mean function of $Y(t)$ in (9.1) is

$$M(t) = E[Y(t)] = E[TC(t) + e(t)] = P(t),$$

since the errors are assumed to have mean zero. Thus a τ-step ahead forecast $\hat{Y}(t+\tau)$ of the mean of $Y(t+\tau)$ is

$$\hat{Y}(t+\tau) = \hat{M}(t+\tau) = \hat{P}(t+\tau)$$

where

$$\hat{P}(t+\tau) = \sum_{i=0}^{q} \hat{s}_i(t)\tau^i/i! \qquad \tau = 1, 2, \cdots \tag{10.11}$$

and where the vector whose components are $\hat{s}_0(t)$, $\hat{s}_1(t)$, $\cdots$, $\hat{s}_q(t)$ is the solution of the equation system (10.3).

For example, when the degree of the polynomial representing the trend-cyclical component is $q = 1$ and we deal with a linear trend,

we have

$$\hat{Y}(t+\tau) = \hat{s}_0(t) + \hat{s}_1(t)\,\tau, \tag{10.12}$$

where $\hat{s}_0(t)$ and $\hat{s}_1(t)$ are computed from the first and second order smoothed values as given by (10.9) and (10.10). Specifically, the $\tau = 1$ step ahead forecast of the expected value of $Y(t)$ is

$$\hat{Y}(t+1) = \hat{s}_0(t) + \hat{s}_1(t) = (2+\tfrac{\alpha}{\beta})S^{(1)}(t) - (1+\tfrac{\alpha}{\beta})S^{(2)}(t). \tag{10.13}$$

To state a more general result for one-step ahead forecasts, we need the concept of the n^{th} order (backward) difference of a time series $Y(t)$. Suppose we define the first difference of $Y(t)$ as

$$\delta Y(t) = Y(t) - Y(t-1).$$

The second order difference can be written

$$\begin{aligned}\delta^2 Y(t) &= \delta Y(t) - \delta Y(t-1) \\ &= Y(t) - 2Y(t-1) + Y(t-2)\end{aligned}$$

and, finally, the n^{th} order difference is

$$\delta^n(t) = \sum_{i=0}^{n} (-1)^i \binom{n}{i} Y(t-i). \tag{10.14}$$

We extend this concept by understanding that $\delta^0 Y(t) = Y(t)$. We now state the following optimality property for the $\tau = 1$ step ahead forecast of the mean of a time series based on exponential smoothing: for

any time series whose n^{th} order difference can be represented as a moving average of order n, n^{th} order exponential smoothing will provide the minimum mean squared error, one-step ahead forecast of the mean.

For the components model

$$Y(t-j) = P(t-j) + e(t-j)$$

where the e(t-j) are uncorrelated with mean zero and have variance proportional to β^{-j} we developed the τ-step ahead model (10.11) through the application of discounted least squares. Thus for $\tau = 1$ we get

$$\hat{Y}(t+1) = \hat{s}_0(t) + \hat{s}_1(t) + \frac{1}{2!}\hat{s}_2(t) + \cdots + \frac{1}{q!}\hat{s}_q(t) \tag{10.15}$$

as the one-step ahead forecast of the mean of Y(t). We now raise the question, is (10.15) an optimal predictor, i.e., is this a minimum mean squared error one-step ahead forecast? The answer to this question depends on whether or not the $(q+1)^{st}$ difference of Y(t) can be written as a moving average process.

Since the time series Y(t) is represented as

$$Y(t-j) = P(t-j) + e(t-j),$$

upon differencing the time series q + 1 times we obtain

$$\delta^{q+1}Y(t-j) = \delta^{q+1}P(t-j) + \delta^{q+1}e(t-j).$$

But $\delta^{q+1}P(t-j) = 0$ since $P(t)$ is a polynomial of degree q and therefore

$$\delta^{q+1}Y(t-j) = \delta^{q+1}e(t-j).$$

We also have from (10.14)

$$\delta^{q+1}e(t-j) = \sum_{i=0}^{q+1} (-1)^i \binom{q+1}{i} e(t-j-i).$$

However, the $\delta^{q+1}e(t-j)$ are not yet expressed as a moving average process because the $e(t-j-i)$ are not white noise. But recalling (9.7) and (9.8), we know that

$$\varepsilon(t-j) = \frac{e(t-j)}{\sigma(t-j)}$$

or

$$e(t-j) = \sigma(t-j)\,\varepsilon(t-j) = \beta^j \varepsilon(t-j)/\sigma^2$$

where σ^2 is a constant of proportionality. Therefore

$$\begin{aligned}
\delta^{q+1}e(t-j) &= \sum_{i=0}^{q+1} (-1)^i \binom{q+1}{i} e(t-j-i) \\
&= \sum_{i=0}^{q+1} (-1)^i \binom{q+1}{i} \beta^i \varepsilon(t-j-i)/\sigma^2 \\
&= \sum_{i=0}^{q+1} (-\beta)^i \binom{q+1}{i} \varepsilon(t-j-i)/\sigma^2 \\
&= \sum_{i=0}^{q+1} b_i \varepsilon(t-j-i)
\end{aligned}$$

where

$$b_i = (-\beta)^i \binom{q+1}{i} / \sigma^2 \qquad i = 0, 1, \cdots, q+1,$$

which expresses $\delta^{q+1} e(t-j)$ as a moving average process of order $q+1$. This establishes (10.15) as an optimal predictor for the one-step ahead forecast.

11. A USEFUL ALTERNATE FORM OF THE SECOND ORDER EXPONENTIAL SMOOTHING MODEL

We recall from the previous section that the second order exponential smoothing forecasting model for the τ-step ahead forecast is

$$\hat{Y}(t+\tau) = \hat{s}_0(t) + \hat{s}_1(t)\,\tau \tag{10.12}$$

in which

$$\hat{s}_0(t) = 2S^{(1)}(t) - S^{(2)}(t) \tag{10.9}$$

and

$$\hat{s}_1(t) = \frac{\alpha}{\beta}(S^{(1)}(t) - S^{(2)}(t)). \tag{10.10}$$

It is convenient for subsequent discussions to express the estimates $\hat{s}_0(t)$ and $\hat{s}_1(t)$ in an alternate form which reflects a recursive relationship between these estimates,

$$\hat{s}_0(t) = \alpha_0 Y(t) + \beta_0[\hat{s}_0(t-1) + \hat{s}_1(t-1)] \tag{11.1}$$

where α_0 denotes the same smoothing constant α involved in (10.9),

(10. 11) and (10. 12), and where

$$\hat{s}_1(t) = \alpha_1(\hat{s}_0(t) - \hat{s}_0(t-1)) + \beta_1\hat{s}_1(t-1) \tag{11.2}$$

where

$$\alpha_1 = \frac{\alpha_0}{1+\beta_0} \quad \text{and} \quad \beta_1 = \frac{\hat{2}\beta_0}{1+\beta_0} \tag{11.3}$$

The relationships (11. 1) and (11. 2) with α_1 and β_1 given by (11. 3) can be obtained from (10. 9), (10. 10), and (10. 12) directly, together with the definitions of the first and second order smoothed values (10. 7). This direct verification of (11. 1) and (11. 2) is somewhat lengthy and the details involved uninteresting for our purposes so they will not be given here. It should be observed that to obtain the same forecast (10. 12) by using (10. 9)) and (10. 10) or by using (11. 1) and (11. 2) for estimates $\hat{s}_0(t)$ and $\hat{s}_1(t)$, it is necessary to choose the smoothing constants α, α_0, and α_1 such that $\alpha_0 = \alpha$ and (11. 3) is satisfied. Moreover, if one chooses these three smoothing constants differently, the forecast based on (11. 1) and (11. 2) will differ from the second order exponential smoothing forecast based on (10. 9) and (10. 10), the latter being obtained from minimizing the discounted sum of squares criterion.

Equation (11. 1) has a natural interpretation which is not apparent from (10. 9). Suppose we denote the $\tau = 1$ step ahead forecast of

the expected value of Y(t) made at time t-1 by $\hat{Y}(t; t-1)$; then

$$\hat{Y}(t; t-1) = [\hat{Y}(t-1+\tau)]$$
$$\tau \quad \tau = 1$$

$$= [\hat{s}_0(t-1) + \hat{s}_1(t-1)\tau]$$
$$\tau = 1 \tau = 1 \tag{11.4}$$

$$= \hat{s}_0(t-1) + \hat{s}_1(t-1).$$

Now the forecast of the expected value of Y(t) made at time t is given by (10.12) with $\tau = 0$,

$$\hat{Y}(t) = [\hat{Y}(t+\tau)]$$
$$\tau = \tau 0 = 0$$

$$= [\hat{s}_0(t) + \hat{s}_1(t)\tau]$$
$$\tau = 0 = 0$$

$$= \hat{s}_0(t).$$

Thus (11.1) can be expressed in an alternate form

$$\hat{Y}(t) = \alpha_0 Y(t) + \beta_0 \hat{Y}(t; t-1) \tag{11.5}$$

which states that the estimate of the expected value of Y(t) made at time t is obtained by simply averaging the current observed value of Y(t) and its forecast value from the immediately preceding period.

Equation (11.2) can be interpreted as follows. We have interpreted $\hat{s}_0(t)$ as an estimate of the expected value of Y(t) at time t;

similarly $\hat{s}_0(t-1)$ is an estimate of the expected value of $Y(t-1)$ made at time $t-1$. Consequently, if there is a linear trend in the expected value of the time series, then the difference $\hat{s}_0(t) - \hat{s}_0(t-1)$ between the two estimates represents the most recent assessment of the slope of the linear trend line that is available on the basis of the current observation $Y(t)$ of the time series. Now (11.2) states that the estimate of the slope of the linear trend component made at time t is merely an average of its most recent assessment $\hat{s}_0(t) - \hat{s}_0(t-1)$ and its previous estimated smoothed value $\hat{s}_1(t-1)$.

12. EXPONENTIAL SMOOTHING MODELS FOR SEASONAL TIME SERIES

Consider a mixed components model of the form

$$Y(t) = TC(t)S(t) + e(t) \tag{12.1}$$

in which the trend-cyclical and seasonal components appear in a multiplicative fashion, and the error or disturbance is additive. It is assumed that the seasonal component $S(t)$ is a periodic function having period ℓ, so that $S(t) = S(t+\ell)$ for all t with ℓ being the smallest positive number for which this property holds.

If $S(t)$ is known, the time series can be <u>deseasonalized</u>, obtaining

$$\frac{Y(t)}{S(t)} = TC(t) + \frac{e(t)}{S(t)} \tag{12.2}$$

or

$$Y^*(t) = TC(t) + e^*(t), \tag{12.3}$$

and one could return to the models of Sections 10 or 11.

It should be emphasized that in using the exponential smoothing models developed earlier it was assumed that seasonal influences were absent (see (9.1) for example). In practice, if one uses these nonseasonal models one should be dealing with data which do not have seasonal influences or one should deseasonalize the data prior to using these nonseasonal exponential smoothing forecasting models.

Extensions of these nonseasonal exponential smoothing models have been developed for some simple cases which permit one to deal with seasonal influences within the model itself. These methods are ad hoc and their statistical and optimality properties from a forecasting point of view are largely unknown at present.

We now develop one such extension and begin by considering the deseasonalized representation of the time series (12.3) in which we assume that the trend-cyclical component is given by a polynomial of degree 1 as

$$TC(t+\tau) = s_0(t) + s_1(t)\tau = P(t+\tau) \tag{12.4}$$

which is a special case of the more general trend-cycle polynomial (10.1) in which $j = -\tau$ and $q = 1$.

In general, were the errors in (12.1) to have mean zero, the expected value of Y(t) for τ periods ahead is given by

$$M(t+\tau) = E[TC(t+\tau)S(t+\tau)) + e(t+\tau)] = TC(t+\tau)S(t+\tau).$$

Hence a τ-step ahead forecast $\hat{Y}(t+\tau)$ of Y(t) would be

$$\hat{Y}(t+\tau) = \hat{TC}(t+\tau)S(t+\tau). \tag{12.5}$$

Now when the trend-cyclical component is given by (12.4), second order exponential smoothing applied to the deseasonalized time series (12.3) provides an estimate of $TC(t+\tau)$

$$\hat{TC}(t+\tau) = \hat{s}_0(t) + \hat{s}_1(t)\tau \tag{12.6}$$

where

$$\hat{s}_0(t) = \alpha_0 Y^*(t) + \beta_0(\hat{s}_0(t-1) + \hat{s}_1(t-1)) \tag{12.7}$$

$$= \alpha_0 \left[\frac{Y(t)}{S(t)}\right] + \beta_0(\hat{s}_0(t-1) + \hat{s}_1(t-1)$$

and

$$\hat{s}_1(t) = \alpha_1(\hat{s}_0(t) - \hat{s}_0(t-1)) + \beta_1\hat{s}(t-1). \tag{12.8}$$

The last two equations correspond to the alternate form of the smoothing model as given in (11.1) and (11.2) except that these are applied to the deseasonalized time series.

If the seasonal component $S(t+\tau)$ were known, (12.7) and (12.8) would be used in (12.6) to forecast $\mathbb{E}[(Y(t+\tau)]$ from (12.5). Typically, however, the seasonal component is unknown and it must be estimated as well. One way to proceed is as follows. Return to the components model (12.1) and assume that TC(t) is known. If we divide both sides of (12.1) by TC(t), we get

$$\frac{Y(t)}{TC(t)} = S(t) + \frac{e(t)}{TC(t)} \tag{12.9}$$

or

$$Y^{**}(t) = S(t) + e^{**}(t). \tag{12.10}$$

The trend-cyclical component of the time series $Y^{**}(t)$ is clearly S(t); thus in order to apply first order exponential smoothing to develop an estimate of S(t) we must assume that S(t) is a constant seasonal factor. Using this interpretation of (12.10) and recalling that $S(t) = S(t-\ell)$, the estimate obtained is

$$\begin{aligned} \hat{S}(t) &= \alpha_s Y^{**}(t) + \beta_s \hat{S}(t-\ell) \\ &= \alpha_s \left[\frac{Y(t)}{TC(t)}\right] + \beta_s \hat{S}(t-\ell). \end{aligned} \tag{12.11}$$

For the linear trend model (12.11) becomes

$$\hat{S}(t) = \alpha_s \left[\frac{Y(t)}{s_0(t)}\right] + \beta_s \hat{S}(t-\ell) \tag{12.12}$$

because $TC(t+\tau) = s_0(t) + s_1(t)\tau$ and therefore $TC(t) = s_0(t)$.

In cases where the linear trend component is unknown, the value of $s_0(t)$ appearing in (12.12) must itself be estimated. If such an estimate is denoted $\hat{\hat{s}}_0(t)$, then an estimate of the smoothed estimate $\hat{S}(t)$ of the seasonal component $S(t)$ would be obtained by replacing $s_0(t)$ by its estimate $\hat{\hat{s}}_0(t)$ in (12.12) and using

$$\hat{\hat{S}}(t) = \alpha_s \left[\frac{Y(t)}{\hat{\hat{s}}_0(t)}\right] + \beta_s \hat{\hat{S}}(t-\ell). \tag{12.13}$$

On the other hand, if $S(t)$ were known, an estimate of $s_0(t)$ would be provided by (12.7). Since the seasonal component is unknown, however, but its period is assumed to be ℓ, its last previous estimate $\hat{\hat{S}}(t-\ell)$ could be used to provide an estimate of its current value $S(t)$. In turn, an estimate of the smoothed estimate $\hat{s}_0(t)$ of the intercept term of the linear trend is provided by using $\hat{\hat{S}}(t-\ell)$ in place of $S(t)$ in (12.7) as

$$\hat{\hat{s}}_0(t) = \alpha_0 \left[\frac{Y(t)}{\hat{\hat{S}}(t-\ell)}\right] + \beta_0[\hat{\hat{s}}_0(t-1) + \hat{\hat{s}}_1(t-1)]. \tag{12.14}$$

Finally, we use this estimate $\hat{\hat{s}}_0(t)$ in (12.8) to obtain in turn the estimate

$$\hat{\hat{s}}_1(t) = \alpha_1[\hat{\hat{s}}_0(t) - \hat{\hat{s}}_0(t-1)] + \beta_1 \hat{\hat{s}}_1(t-1). \tag{12.15}$$

The forecast value $\hat{\hat{Y}}(t+\tau)$ would then be given by

$$\hat{\hat{Y}}(t+\tau) = \hat{\hat{TC}}(t+\tau)\hat{\hat{S}}(t+\tau-\ell) \tag{12.16}$$

where

$$\hat{\hat{TC}}(t+\tau) = \hat{\hat{s}}_0(t) + \hat{\hat{s}}_1(t)\tau. \tag{12.17}$$

The equations (12.13), (12.14) and (12.15), together with (12.16) and (12.17), are called Winters' seasonal exponential smoothing model for a linear trend with a multiplicative seasonal component [23].

It should be noted that in order to use this model one must choose the smoothing constants α_s, α_0, and α_1 in equations (12.13), (12.14), and (12.15) and, moreover, one must know the periodicity of the seasonal component as expressed in the choice of the period ℓ. Thus, four parameters must be chosen from an analysis of the data or by other means in order to apply Winters' multiplicative exponential smoothing model.

Given the period ℓ of the seasonal component, Winters has provided a heuristic procedure for choosing α_s, α_0, and α_1 which consists of making various selections of the three smoothing constants over a grid of possible values. Then for each selection of α_s, α_0, and α_1, one-step ahead forecasts are made and the root mean squared error is calculated for each selection. One chooses that selection of α_s, α_0, and α_1 which is associated with the smallest of the calculated

root mean squared errors. This is an attempt to achieve a selection of the three smoothing constants which minimizes the mean squared error of the one-step ahead forecasts.

Three aspects of Winters' method should be noted. First, it must be assumed that the grid search procedure produces a selection of the smoothing constants α_s, α_0, and α_1 which is reasonably close to that selection which minimizes the mean squared errors of the forecast over the sequence of observations. Second, it must also be assumed that this choice based on the past data remains a good choice for making forecasts for future time periods; and third, that the period ℓ of the seasonal influence is known exactly.

In Winters' model the seasonal influence is represented in the choice of the period ℓ and the resulting seasonals are given as a system of adjustments that are to be applied to the time series in accordance with the equations defining the model. In other words, the seasonal influence is not described explicitly as a trigonometric function.

Harrison [14] has developed an alternative approach in which the seasonal influence is fitted by means of a trigonometric function and then the seasonal estimates are smoothed and used in forecasting.

The multiplicative seasonal exponential smoothing approach has been recast as an additive seasonal model by McClain [18] and McClain and Thomas [19].

Burman [5] has used trigonometric functions to estimate seasonal influences in an additive components model.

In all the exponential smoothing models above, the smoothing values α are constants and must be chosen by the user of the model. It is natural to consider a more general exponential smoothing model in which α is not a constant but depends on time. Such a model, which can be called a dynamic exponential smoothing model, has been developed by Trigg and Leach [22] and has been used by Dunn, Williams, and Spivey [8].

13. BOX-JENKINS FORECASTING PROCEDURES

In using the exponential smoothing models in preceding sections one oftentimes does not attempt to analyze the underlying probability structure of the time series by methods of statistical inference. The smoothing constant α, although related to the probability structure of the additive error terms of the underlying components model, is often chosen by the user without reference to more objective considerations of statistical inference. Moreover, as we have also emphasized, these models require one to assume that the trend-cyclical component is a polynomial of known degree q. The simple recursive updating relationships in these models, however, allow them to be applied conveniently in an almost mechanical fashion and this simplicity has made

these models widely attractive in situations where many time series must be forecast repeatedly on a routine basis as in parts inventory systems.

In contrast to the features above, Box-Jenkins [3] have developed forecasting procedures which require one to analyze the underlying probability structure of the time series in considerably more detail and methods of statistical inference are employed systematically to estimate properties of this underlying structure.

Secondly, these procedures are more flexible than exponential smoothing models with respect to the specification of the underlying trend-cyclical component of the time series. For example, in exponential smoothing we must assume that this is a polynomial of degree q, whereas Box-Jenkins procedures provide various transformations which can be performed on the data and which accommodate more varied trend-cyclical components. Thus if the data appear to have an exponential trend, one would take logarithms of the data and proceed with the analysis. If the logarithms of the data themselves appear to have a linear trend, one could difference the logarithms. In short, without having to specify the trend-cyclical component explicitly and then estimate it, the Box-Jenkins procedure provides a flexible system of simple transformations which are applied to the data until finally an autoregressive moving average process of order (p, q) appears to result.

Thirdly, unlike exponential smoothing models which have simple updating features, Box-Jenkins procedures require more elaborate computer algorithms for obtaining estimates of the parameters of the ARMA process, and there is no simple recursive method for updating these estimates as additional observations become available.

We now proceed to a discussion of the Box-Jenkins procedure. Suppose that after a time series $Y(t)$ is differenced d times we arrive at a time series $Z(t) = \delta^d Y(t)$, where it is known that $Z(t)$ is an ARMA(p, q) process of the form (8. 1), namely

$$(13.1) \qquad Z(t) - \phi_1 Z(t-1) - \cdots - \phi_p Z(t-p) = \varepsilon(t) - \theta_1 \varepsilon(t-1) - \cdots - \theta_q \varepsilon(t-q) \qquad - \theta_q \varepsilon(t-q).$$

We assume that the $\varepsilon(t)$ are white noise and that all the roots of the associated polynomial equation lie outside the unit circle, so that $Z(t)$ is a covariance stationary process. When $d > 0$, $Y(t)$ is called an integrated autoregressive moving average process and is denoted by ARIMA(p, d, q).

When the expected value of the original time series $Y(t)$ is a constant $\mu \neq 0$ we will consider

$$\tilde{Y}(t) = Y(t) - \mu$$

instead of $Y(t)$, and if $\tilde{Y}(t)$ is an ARMA(p, q) process with mean zero we say that $Y(t)$ is an ARMA process with mean μ. In addition we generalize (13. 1) slightly by permitting the $\varepsilon(t)$ to have constant variance σ_ε^2 rather than unit variance as formerly. These minor extensions concerning the mean and variance, unnecessary in earlier discussions, are now made to permit us to consider a slightly wider class of economic time series.

For purposes of forecasting the time series $Z(t)$, Box-Jenkins suggest using of (13. 1) which requires the estimation of p+q parameters together with the variance σ_ε^2 and they refer to (13. 1) as a parsimonious representation.

Obtaining optimal or minimum mean squared error τ-step ahead forecasts for models of the form (13. 1) is simplified because $Z(t)$ and $\varepsilon(t)$ appear in (13. 1) in essentially linear fashion. To see why this is the case, let $\hat{Z}(t+\tau)$ denote a forecast of the expected value of $Z(t+\tau)$ based on a linear representation involving the current and all previous observations as

$$\hat{Z}(t+\tau) = \Pi_0 Z(t) + \Pi_1 Z(t-1) + \Pi_2 Z(t-2) + \cdots \tag{13.2}$$

$$= \sum_{j=0}^{\infty} \Pi_j Z(t-j) .$$

Since, however, $Z(t)$ can be represented as a moving average process of infinite order,

$$Z(t) = \sum_{i=0}^{\infty} \psi_i \varepsilon(t-i) \tag{13.3}$$

with $\psi_0 = 1$, we find that the forecast (13.2) can be expressed as a linear combination of only the error terms associated with the moving average part of the process,

$$\hat{Z}(t+\tau) = \sum_{j=0}^{\infty} \Pi_j Z(t-j)$$

$$= \sum_{j=0}^{\infty} \Pi_j \sum_{i=0}^{\infty} \psi_i \varepsilon(t-i-j) \tag{13.4}$$

$$= \sum_{k=0}^{\infty} \Pi_k^* \varepsilon(t-k)$$

where

$$\Pi_k^* = (\Pi_0 + \Pi_1 + \cdots + \Pi_k)\psi_k .$$

Therefore the difference between the actual value $Z(t+\tau)$ at time τ and its forecast value $\hat{Z}(t+\tau)$ at time τ can be expressed as

$$Z(t+\tau) - \hat{Z}(t+\tau) = \sum_{i=0}^{\tau-1} \psi_i \varepsilon(t+\tau-i) + \sum_{k=0}^{\infty} (\psi_{\tau+k} - \Pi_k^*) \varepsilon(t-k) \tag{13.6}$$

using (13.3) and (13.4). Consequently,

$$E[(Z(t+\tau) - \hat{Z}(t+\tau))^2] = \sigma_\varepsilon^2 \sum_{i=0}^{\tau-1} \psi_i^2 + \sigma_\varepsilon^2 \sum_{k=0}^{\infty} (\psi_{\tau+k} - \Pi_k^*) .$$

From the decomposition of the sum of squares above we see that

$\hat{Z}(t+\tau)$ will be the minimum mean squared error or optimal forecast if and only if the coefficients Π_k^* are chosen so as to be identical to the corresponding coefficients $\psi_{\tau+k}$ which appear in the representation of $Z(t+\tau)$ as a moving average process of infinite order (13.3). Therefore (13.4) can be expressed as

$$\hat{Z}(t+\tau) = \sum_{k=0}^{\infty} \psi_{\tau+k} \varepsilon(t-k), \tag{13.7}$$

and the mean squared error associated with this forecast is

$$E[(Z(t+\tau) - \hat{Z}(t+\tau))^2] = \sigma_\varepsilon^2 \sum_{i=0}^{\tau-1} \psi_i^2 ,$$

the mean squared error in this case, of course, being identical to the variance of the forecast error.

An alternate expression for this optimal forecast is obtained by returning to the ARMA(p, q) process which has (13.7) as its infinite order moving average representation. To develop such an expression we first observe that $\hat{Z}(t+\tau)$ as expressed in (13.7) is simply the conditional expectation of $Z(t+\tau)$ given $Z(t)$, $Z(t-1)$, $\cdots$, or equivalently the conditional expectation of $Z(t+\tau)$ given $\varepsilon(t)$, $\varepsilon(t-1)$, $\cdots$. In other words, if we calculate the conditional expectation $Z(t+\tau)$ given $\varepsilon(t)$, $\varepsilon(t-1)$, $\cdots$, we see that

$$(13.8)\quad E[Z(t+\tau)\,|\,\varepsilon(t),\,\varepsilon(t-1)\cdots] = E\Big[\sum_{i=0}^{\infty}\psi_i\,\varepsilon(t+\tau-i)\,\Big|\,\varepsilon(t),\,\varepsilon(t-1),\cdots\Big]$$

$$= E\Big[\sum_{i=0}^{\tau-1}\psi_i\,\varepsilon(t+\tau-i)\,\Big|\,\varepsilon(t),\,\varepsilon(t-1),\cdots\Big]$$

$$+ E\Big[\sum_{i=\tau}^{\infty}\psi_i\,\varepsilon(t+\tau-i)\,\Big|\,\varepsilon(t),\,\varepsilon(t-1),\cdots\Big].$$

When the successive future errors $\varepsilon(t+1),\ \cdots,\ \varepsilon(t+\tau)$ are independent of the errors $\varepsilon(t),\ \varepsilon(t-1),\ \cdots$ of the current and previous periods, then

$$E\Big[\sum_{i=0}^{\tau-1}\psi_i\,\varepsilon(t+\tau-i)\,\Big|\,\varepsilon(t),\,\varepsilon(t-1),\cdots\Big] = E\Big[\sum_{i=0}^{\tau-1}\psi_i\,\varepsilon(t+\tau-i)\Big].$$

Also, since the mean of $\varepsilon(t)$ is zero,

$$E\Big[\sum_{i=0}^{\tau-1}\psi_i\,\varepsilon(t+\tau-i)\Big] = \sum_{i=0}^{\tau-1}\psi_i\,E[\varepsilon(t+\tau-i)] = 0$$

we obtain

$$(13.9)\quad E[Z(t+\tau)\,|\,\varepsilon(t),\,\varepsilon(t-1),\cdots] = E\Big[\sum_{k=0}^{\infty}\psi_{\tau+k}\,\varepsilon(t-k)\,\Big|\,\varepsilon(t),\,\varepsilon(t-1),\cdots\Big]$$

$$= \sum_{k=0}^{\infty}\psi_{\tau+k}\,\varepsilon(t-k)$$

showing as asserted that

$$\hat{Z}(t+\tau) = E[Z(t+\tau) \mid \varepsilon(t), \varepsilon(t-1), \cdots]. \tag{13.10}$$

Suppose we take (13.9) and reexpress it as an ARMA(p, q) process; then we obtain

$$Z^*(t+\tau; t) - \phi_1 Z^*(t+\tau-1; t) - \cdots - \phi_p Z^*(t+\tau-p; t) = \tag{13.11}$$

$$\varepsilon^*(t+\tau; t) - \theta_1 \varepsilon^*(t+\tau-1; t) - \cdots - \theta_q \varepsilon^*(t+\tau-q; t)$$

where the $Z^*(t+\tau; t)$, $\varepsilon^*(t+\tau; t)$, etc., denote random variables which represent the conditional expectation involved, namely

$$Z^*(t_2; t_1) = E[Z(t_2) \mid \varepsilon(t_1), \varepsilon(t_1-1), \cdots]$$

$$\varepsilon^*(t_2, t_1) = E[\varepsilon(t_2) \mid \varepsilon(t_1), \varepsilon(t_1-1), \cdots].$$

Therefore the forecast $\hat{Z}(t+\tau)$ is expressed in the form of an ARMA(p, q) process as

$$\hat{Z}(t+\tau) = \sum_{i=1}^{p} \phi_i Z^*(t+\tau-i; t) + \sum_{j=1}^{q} \theta_j \varepsilon^*(t+\tau-j; t) \tag{13.12}$$

$$+ \varepsilon^*(t+\tau; t)$$

Thus to generate the optimal τ-step ahead forecast the parameters $\phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q$ as well as the various conditional expectations appearing in (13.12) must be known or be estimated from the data.

The Box-Jenkins procedure provides two ways to calculate the conditional expectations, one is to use the unconditional expectations and the other is to utilize the "back-forecasting" scheme. These methods are discussed in Chapter 6 of [3].

When the finite dimensional distributions of the $\varepsilon(t)$ are multivariate normal, maximum likelihood estimators of the parameters $\phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q$ and σ_ε^2 can be obtained as solutions to a system of simultaneous non-linear equations (see Chapter 7 of [3]). Other estimators are also derived based upon the minimization of sums of squares related to the quadratic form associated with the multivariate normal distribution. These approximations concentrate on minimizing the quadratic form rather than the likelihood function which would be involved in the exact maximum likelihood procedure.

Each of the preceding forecasting and estimation procedures assumes that we know the order of differences d of the time series of observations and the orders p and q of the autoregressive and moving average processes. Box and Jenkins have also provided identification and diagnostic procedures to assist the user to determine the numbers p, d, and q by classical methods of statistical inference that involve the Yule-Walker equations and sample estimates of the autocorrelations and partial autocorrelations of the data as well as the estimated forecast errors.

Determination of p, d, and q is a crucial part of the Box-Jenkins procedure and can lead to different forecasting models being generated from the same data by different users. Since these methods are somewhat unsatisfactory, others in recent research have suggested alternate approaches to the identification and diagnostic procedures of Box and Jenkins (see Parzen [21]).

BIBLIOGRAPHY

1. Ash, R. B. Real Analysis and Probability. New York: Academic Press, 1972.

2. Blackman, R. B., and Tukey, J. The Measurement of Power Spectra. New York: Dover Publications, 1959.

3. Box, G. E. P., and Jenkins, G. Time Series Analysis, Forecasting and Control. San Francisco: Holden-Day, 1970.

4. Brown, Robert G. Smoothing, Forecasting and Prediction of Discrete Time Series. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1962.

5. Burman, J. P. "Moving Seasonal Adjustment of Economic Time Series." Journal of the Royal Statistical Society, Series A, 128 (1965): 534-558.

6. Cogger, K. O. "Extensions of the Fundamental Theorem of Exponential Smoothing." Management Science 19 (Jan. 1973): 547-554.

7. Doob, J. L. Stochastic Processes. New York: John Wiley and Sons, 1953.

8. Dunn, D. M., Williams, W. H., and Spivey, W. A. "Analysis and Prediction of Telephone Demand in Local Geographical Areas." The Bell Journal of Economics and Management Science 2 (Autumn, 1971): 561-576.

9. Feller, W. An Introduction to Probability Theory and Its Applications, vol. I. New York: John Wiley and Sons, 1957.

10. Feller, W. An Introduction to Probability Theory and Its Applications, vol. II. New York: John Wiley and Sons, 1966.

11. Granger, C. W. J. and Hatanaka, M. Spectral Analysis of Economic Time Series. Princeton: Princeton University Press, 1964.

12. Grenander, V., and Rosenblatt, M. Statistical Analysis of Stationary Time Series. New York: John Wiley and Sons, 1957.

13. Hannan, E. J. Time Series Analysis. New York: John Wiley and Sons, 1960.

14. Harrison, P.J. "Short-Term Sales Forecasting." Applied Statistics 14 (1965): 102-134.

15. Jenkins, G., and Watts, D.G. Spectral Analysis and Its Applications. San Francisco: Holden-Day, 1968.

16. Kendall, M.G. Time-Series. New York: Hafner Publishing Co., 1973.

17. Kendall, M.G., and Stuart, A. The Advanced Theory of Statistics, vol. 3. New York: Hafner Publishing Co., 1968.

18. McClain, J.O. "Dynamics of Exponential Smoothing with Trend and Seasonal Terms." Management Science 20 (May, 1974): 1300-1304.

19. McClain, J.O., and Thomas, L.J. "Response-Variance Tradeoffs in Adaptive Forecasting." Operations Research 21 (March-April 1973), 554-568.

20. Nelson, C.R. Applied Time Series Analysis. San Francisco: Holden-Day, 1973.

21. Parzen, E. "Some Solutions to the Time Series Modeling and Prediction Problem." Technical Report No. 5, Department of Computer Science, State University of New York at Buffalo, Feb. 1974.

22. Trigg, D.W., and Leach, A.G., "Exponential Smoothing with an Adaptive Response Rate." Operational Research Quarterly 18 (1964): 53-59.

23. Winters, P.R. "Forecasting Sales by Exponentially Weighted Moving Averages." Management Science 6 (1960): 324-342.

24. Yaglom, A.M. An Introduction to the Theory of Stationary Random Functions. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1962.