

**TIMES**  
**Box-Jenkins Forecasting System**

**Reference Manual**

**Volume I**  
**TECHNICAL BACKGROUND**

**Joseph George Caldwell, PhD**

Revised March 1971  
(Reformatted September 2006)

© 1971, 2006 Joseph George Caldwell. All Rights Reserved.

Posted at Internet websites <http://www.foundationwebsite.org> and <http://www.foundation.bw>

**Note:** The document *The Box-Jenkins Forecasting Technique*, posted at <http://www.foundationwebsite.org/BoxJenkins.htm>, presents a nontechnical description of the Box-Jenkins methodology. For a technical description of the Box-Jenkins approach, see the document, *TIMES Box-Jenkins Forecasting System*, posted at <http://www.foundationwebsite.org/TIMESVol1TechnicalBackground.pdf>. A set of briefing slides describing mathematical forecasting using the Box-Jenkins methodology is posted at [http://www.MathematicalForecasting\\_Box-Jenkins.pdf](http://www.MathematicalForecasting_Box-Jenkins.pdf). A computer program that can be used to develop a broad class of Box-Jenkins models is posted at the Foundation website, <http://www.foundationwebsite.org> (6 February 2009).

## Table of Contents

I. SOME THEORETICAL ASPECTS OF TIME SERIES ANALYSIS .....	1
A. INTRODUCTION .....	1
1. The Need for Time Series Models .....	1
2. General Considerations on Stochastic Model Construction .....	1
3. Limitations Associated with Stochastic Model Construction .....	3
B. NATURE OF THE TIME SERIES MODEL .....	4
1. The Mixed Autoregressive Moving Average Process .....	4
2. Multiplicative Seasonal Components .....	5
3. Deterministic Components .....	6
C. CHARACTERISTICS OF THE TIME SERIES MODEL .....	6
1. Stationarity .....	6
2. The Autocorrelation Function .....	7
3. The Periodogram and the Spectrum .....	8
a. The Periodogram .....	8
b. The Spectrum .....	8
4. Behavior of a Process as a Function of its Model Parameters .....	10
D. STATISTICAL ANALYSIS OF THE TIME SERIES MODEL .....	11
1. Tentative Model Selection .....	11
a. Differencing for Stationarity .....	11
b. The autocorrelation Function and Partial Autocorrelation Function .....	12
(1) Pure Moving Average Process .....	12
(2) Pure Autoregressive Process .....	12
(3) Mixed Autoregressive Moving Average Model .....	12
c. The Periodogram and the Spectrum .....	12
2. Estimation .....	13
a. Least-Squares Estimates .....	13
b. Maximum Likelihood Estimates .....	13
3. Diagnostic Checking .....	13
a. Hypothesis Testing .....	13
b. Significance of Parameters .....	14
c. Analysis of Residuals .....	14
(1) Autocorrelation Analysis .....	14
(2) Periodogram Analysis .....	15
(3) Spectral Analysis .....	15
d. Over-fitting .....	15
e. Independent Estimates from Different Data .....	15
E. SIMULATION AND FORECASTING .....	16
1. Simulation .....	16
2. Forecasting .....	16
a. Derivation of the Least-Squares Forecaster .....	17
b. Computation of the Forecast .....	18
c. Eventual Forecast Function .....	19
II. THE TIME SERIES ANALYSIS PROGRAM <i>TIMES</i> .....	19
A. THE ESTIMATION PROGRAM “ESTIMATE” .....	19
B. THE PROJECTION PROGRAM “PROJECTOR” .....	21

REFERENCES .....	21
APPENDIX A. AUTOCORRELATION ANALYSIS .....	22
APPENDIX B. PERIODOGRAM ANALYSIS .....	24
APPENDIX C. SPECTRAL ANALYSIS .....	26
1. Definition of the Spectrum .....	26
2. Estimates of the Spectrum .....	27
a. Raw Periodogram .....	27
b. Smoothed Periodogram .....	27
c. Weight Functions .....	28
3. White Noise .....	29
4. A Test for Whiteness .....	29
APPENDIX D. LINEAR DIFFERENCE EQUATIONS .....	32
APPENDIX E. LEAST-SQUARES AND MAXIMUM LIKELIHOOD ESTIMATION .....	33
1. Introduction .....	33
2. Least-Squares Estimates .....	33
a. Pure Autoregressive Model .....	33
b. Mixed Autoregressive Moving Average Model .....	34
3. Maximum Likelihood Estimates .....	36
4. The Likelihood Function .....	36
APPENDIX F. TRANSFORMATION OF DATA .....	38
APPENDIX G. EXPONENTIAL SMOOTHING .....	40

## Preface

When Lambda's *TIMES* package was first developed there was little readily available published material describing the theoretical work that was being conducted by Professors G.E.P. Box and G.M. Jenkins. This Technical Background was prepared in order to provide *TIMES* users with the basic concepts of the Box-Jenkins approach. Now that the Box-Jenkins volume (Reference 1) is published and available we strongly encourage the *TIMES* user to consult it for an in-depth presentation of the method. We believe, however, that the present volume constitutes not only a useful introduction to the *TIMES* package, but also a handy summary of that part of the Box-Jenkins book relating to forecasting.

Comments and criticisms on the presentation are most welcome.

Joseph George Caldwell, PhD (Statistics)

## Summary

The computer program *TIMES* has been developed to enable rapid development of models from time series data. This document describes a number of results from the theory of time series analysis, an understanding of which results is considered necessary for proper use of the program as an analysis tool. It is written for the technical analyst who wishes to use the *TIMES* program. Its purpose is twofold: (1) to familiarize the analysts with basic concepts of time series modeling and the capabilities of the *TIMES* program; (2) to serve as a self-contained reference for the analyst to assist his application of *TIMES*. Familiarity with this document is considered necessary for proper application of *TIMES*, in view of the possibility of deriving incorrect or incomplete models through a lack of understanding of the implications of the analysis performed by *TIMES*.

The main text of the document describes the *TIMES* series model, its properties, and certain statistical concepts and results in summary form. Detailed technical descriptions of fundamental statistical analysis techniques mentioned in the text are included in the appendices.

A description of the procedures for using the *TIMES* program is included in the *TIMES Reference Manual*, Volume II.

# I. SOME THEORETICAL ASPECTS OF TIME SERIES ANALYSIS

## A. INTRODUCTION

### 1. The Need for Time Series Models

In many economic and industrial situations, it is necessary to predict, or forecast, the value of a particular quantity over some future time period. Often, data are available describing the past behavior of the process whose future behavior we wish to predict. In order to be able to make accurate forecasts it is desirable to be able to relate the future behavior of the process to its past behavior. In other words, we need a *model* of the process.

It is often the case that we are interested in predicting the behavior of a variable over the near future, based on its behavior over the recent past. In such a situation, we need a good model of the short-term behavior of the process. Depending on the application, we may or may not want the model to involve variables additional to the variable whose value we wish to forecast. In any event, we would like a model that describes the behavior of the variable of interest as accurately as possible, in terms of whatever variables are included in the model. Such a model can be used to simulate, as well as to predict, the behavior of the variable of interest.

Recently, a class of models has been investigated that can efficiently represent the short-term behavior of many of the types of time series that occur in economic and industrial contexts. These models, described by Box and Jenkins (Reference 1), are particularly useful for developing stochastic models which involve solely the single variable we wish to forecast (or simulate). They are also quite useful for situations in which the behavior of the variable to be forecast is modeled in terms of the behavior of other variables as well.

The computations associated with the analysis required to develop a Box-Jenkins model from time series data can be quite extensive. The time series analysis program, *TIMES*, has been developed to perform these computations and assist in the efficient development of these models.

This document summarizes the aspects of time series analysis that are considered most important for effective application of the *TIMES* program. A comprehensive treatment of the subject can be found in Box and Jenkins' book (Reference 1).

### 2. General Considerations on Stochastic Model Construction

In general, a stochastic model describing a process expresses the variable of interest (dependent variable) in terms of other observable variables (independent variables), a random "noise" variable ("error" term), and a number of parameters. The functional form that parametrically relates the dependent variable to the independent variables can be suggested by prior physical considerations, or arise through the course of analysis of the process. The noise variable typically represents the effects of all variables (sources of variation) not explicitly taken into account in the model. Conceivably, a better understanding of the physical process generating the observations would enable construction of a model with more explanatory variables (or a different functional representation), and a "smaller" error term (i.e., a noise variable with smaller variance). An important assumption in the statistical analysis of time series is that the error terms are uncorrelated with the independent variables of the model.

The general approach to statistical model construction is to use whatever theoretical knowledge is available to suggest a functional form for the model. Statistical methods are then used to determine the

number of terms necessary and to estimate values of the parameters of the model. The dynamics of many physical systems can be expressed in terms of a differential equation

$$(1 + c_1D + c_2D^2 + \dots)Y = (1 + d_1D + d_2D^2 + \dots)X$$

where  $Y$  is the output variable and  $X$  is the input variable,  $D$  is the differential operator, and the  $c$ 's and  $d$ 's are constants. Such simple models often describe complex systems adequately, even when the true nature of the system is not understood.

In the discrete case in which observations are taken at equally spaced intervals, the above *differential* equation is replaced by a *difference* equation

$$(1 + c_1\nabla + c_2\nabla^2 + \dots)Y_t = (1 + d_1\nabla + d_2\nabla^2 + \dots)X_t$$

where  $\nabla$  denotes the backward difference operator, defined by  $\nabla Y_t = Y_t - Y_{t-1}$ . Similarly, the complex stochastic behavior of a random process  $\{z_t\}$  can often successfully be described in terms of a difference equation relating  $\{z_t\}$  to a much simpler random process – a “white noise” process,  $\{a_t\}$ , having zero mean, constant variance, and no correlation among its members:

$$(1 + c_1\nabla + c_2\nabla^2 + \dots)z_t = (1 + d_1\nabla + d_2\nabla^2 + \dots)a_t \quad (1)$$

We shall assume that there are but a finite number of  $c$ 's and  $d$ 's. it is convenient here to introduce the backward shift operator  $B$ , defined by  $Bz_t = z_{t-1}$ , and to rewrite the preceding equation in terms of  $B = 1 - \nabla$ :

$$(1 - \phi_1B - \phi_2B^2 - \dots - \phi_pB^p)z_t = (1 - \theta_1B - \theta_2B^2 - \dots - \theta_qB^q)a_t \quad (2)$$

Applying the polynomial operators to  $z_t$  and  $a_t$ , we may write this model as

$$z_t - \phi_1z_{t-1} - \dots - \phi_pz_{t-p} = a_t - \theta_1a_{t-1} - \dots - \theta_qa_{t-q}$$

or

$$z_t = \phi_1z_{t-1} + \dots + \phi_pz_{t-p} + a_t - \theta_1a_{t-1} - \dots - \theta_qa_{t-q} \quad (3)$$

This equation defines the basic model investigated by Box and Jenkins. In words, the current observation is assumed to be represented in terms of a finite linear combination of previous observations, plus a white noise error term associated with the current time period, plus a finite linear combination of the white noise term associated with previous time periods.

It is often the case that more than one model can be determined to describe a process. For example, the models

$$z_t = a_t + \theta a_{t-1} \quad (4)$$

and

$$z_t = a_t + \theta z_{t-1} + \theta^2 z_{t-2} + \dots \quad (5)$$

where  $|\theta| < 1$ , are equivalent. Owing to errors in estimation of the coefficients, a fitted model of form (5) might not be recognized as an alternative form of (4). Clearly the representation (4) is a much more

efficient characterization of the process  $\{z_t\}$ . It is a generally accepted principle of model construction that the model adequately describing the process with the fewest parameters is chosen to represent the process.

Box and Jenkins have succinctly described the iterative procedure by which models are determined:

1. Decide, on the basis of theory and experience, on a useful class of models;
2. Determine, using knowledge of the system and observed data, a tentative subclass of models;
3. Estimate the parameters of the tentative model from the data;
4. Apply diagnostic checks of the statistical adequacy of the fitted model.

If a lack of fit is discovered, attempt to diagnose the cause, modify the subclass of tentative models, and repeat steps (3) and (4).

This document describes the basic characteristics of time series, and the nature and properties of Box-Jenkins time series models. We shall indicate how the “statistics” associated with a time series are related to the Box-Jenkins model of the time series, and describe the diagnostic checks.

Note that a model of a process may involve both dynamic and stochastic components; i.e., the model may include other variables in addition to the white noise variable ( $a_t$ ). Although *TIMES* can analyze certain models which include additional variables, this capability was not a primary motivation in its development. Because the principal application of the *TIMES* program is the single-variable situation, the reader is referred to Box-Jenkins’ book (Reference 1) for the (fairly involved) theory underlying the multiple-variable case.

### 3. Limitations Associated with Stochastic Model Construction

The purpose of the time series analysis program, *TIMES*, is to assist the user in inferring an underlying stochastic model from a set of real-valued observations taken at equally spaced intervals (in time or space). By the term “model” is meant a description of the stochastic process generating the observations in terms of a “small” number of parameters (to be estimated from the observed data), and a “white noise” sequence of random variables having significantly smaller variance than that of the original process. Since, based on past observations with less error variability than the variability of the white noise process itself, deriving a model from a set of time series observations may be regarded as identifying the predictable behavior associated with the series. The remaining unpredictable component of behavior is embodied in the white noise process.

The adequacy of a model as a statistical representation of a process can be checked by testing the model “residuals”, or estimates of the white noise sequence corresponding to the process, for departure from “whiteness”. There are two principal reasons for doing this. First, the parameter values are estimated assuming whiteness of the residuals; these estimates, although biased, and “consistent” and “efficient” if the residuals are white and normally distributed. If the residuals are not white, the estimates are biased, inconsistent, and inefficient. Second, whiteness of the residuals suggests that the predictable behavior of the process which is related to the sampled information has been identified. (That is, there are no more parameters to estimate for the parametric class of model chosen.) It is important to note that it may be possible to derive a much more precise model by including additional variables in the model. Both models could be adequate statistical representations of the process, in terms of their respective variables. Selection of the variables to include in a model is guided by physical considerations or experience; time series analysis is intended to find statistically adequate models in terms of the variables that have been included in the model.

(For example, consider a stochastic process

$$z_t = ax_{t-1} + by_{t-1} + a_t$$

where  $x_t$ ,  $y_t$  and  $a_t$  are uncorrelated white noise processes (the  $a_t$ 's being the "residuals"). If we observe only the  $z_t$  and  $x_t$ , we will determine the model

$$z_t = \widehat{a}x_{t-1} + b_t$$

where the residuals are now the white noise sequence  $b_t = by_{t-1} + a_t$ . Had we also observed  $y_t$  however, we would derive the model

$$z_t = \widehat{a}x_{t-1} + \widehat{b}y_{t-1} + a_t .$$

Both models are adequate representations of the stochastic behavior of the process  $z_t$ . The latter representation, however, is more complete, more precise, and if  $b$  is large and the  $y_t$  are easy to observe, likely to be more useful.)

The estimated residuals corresponding to a model are a random sample, and hence a test of their "whiteness" will be a statistical one. We can hence never be absolutely certain that the residuals are from a true white-noise process. Because of this, there may be a number of alternative models which describe a given stochastic process adequately. We cannot say for certain which is the "best" model. In general, however, one model can be chosen as being the most satisfactory on the basis of its number of parameters, or its residual error variance, or the "whiteness" of its residuals, or perhaps even on the basis of the reasonableness of the functional form of the parameters of the model.

## B. NATURE OF THE TIME SERIES MODEL

### 1. The Mixed Autoregressive Moving Average Process

As discussed in the introduction, the general class of models that has been found to be quite useful for modeling time series has the following form:

$$(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) z_t = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q) a_t \quad (6)$$

where  $\{z_t\}$  is the stochastic process being modeled,  $\{a_t\}$  is a white noise process (i.e., a sequence of uncorrelated random variables having zero mean and constant variance),  $B$  is the backward shift operator, defined by  $Bz_t = z_{t-1}$ , and the  $\phi$ 's and  $\theta$ 's are the (unknown) parameters of the model, to be estimated from a realization of the process (i.e., a sample of successive  $z$ 's) as noted earlier, the above formula (6) is simply a shorthand notation for

$$z_t - \phi_1 z_{t-1} - \dots - \phi_p z_{t-p} = a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q} . \quad (7)$$

We assume that  $z_t$  represents a deviation from a mean value. If this mean value is constant, we may drop this assumption by introducing an additional parameter, into the model:

$$(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) z_t = \theta_0 + (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q) a_t . \quad (8)$$

For simplicity, we shall drop the parameter  $\theta_0$  from the following discussion.

Defining the polynomials

$$\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$$

and

$$\theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q$$

we may write the model (7) as

$$\Phi(B) z_t = \Theta(B) a_t . \quad (9)$$

There are some cases of model (9) that are of special interest. If  $\Theta(B) = 1$ , we may write the model as

$$\Phi(B)z_t = a_t$$

or

$$z_t = \phi_1 z_{t-1} + \dots + \phi_p z_{t-p} + a_t .$$

The model is hence simply a regression model of the most recent  $z_t$  on previous  $z_t$ 's, and the model is called an *autoregressive* process of order  $p$ . If, on the other hand,  $\Phi(B) = 1$ , then the model becomes

$$z_t = \Theta(B)a_t$$

or

$$z_t = a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q} .$$

Thus  $z_t$  is a moving average of the “error terms”, and the process is called a *moving average* process of order  $q$ . The general model (9) is called a *mixed autoregressive moving average* process.

## 2. Multiplicative Seasonal Components

While the general form of the mixed autoregressive moving average model (9) is capable of modeling quite general processes  $z_t$ , there are often special characteristics of a particular process that can be used to justify use of a model that has a more specific structure than (9). Such is the case in modeling processes in which seasonal behavior is inherent. A reasonable model is the following:

$$\Phi_s(B^s)z_t = \Theta_s(B^s)e_t$$

where  $s$  is the seasonal period; that is, the current  $z$  ( $z_t$ ) is related to the corresponding  $z$ 's ( $z_{t-s}, z_{t-2s}, \dots$ ) from previous periods. However, the error terms ( $e_t, e_{t-s}, \dots$ ) included in the model are expected to be correlated with nearby error terms. We hence assume a model

$$\Phi_1(B)e_t = \Theta_1(B)a_t$$

for the  $e_t$ 's, where the  $a_t$ 's are white noise. Combining the above two expressions, we have

$$\Phi_s(B^s)\Phi_1(B)z_t = \Theta_s(B^s)\Theta_1(B)a_t \quad (10)$$

as a reasonable model expressing the  $z_t$ 's in terms of white noise residuals. The model is, of course, a special case of (9) in which  $\Phi = \Phi_s\Phi_1$  and  $\Theta = \Theta_s\Theta_1$ .

Since the autoregressive and moving average polynomials are multiplicative combinations, we call the model a multiplicative seasonal model. In general, if we expect  $c$  seasonal components, of periods  $s_1, s_2, \dots, s_c$ , then our model is

$$\Phi_{s_c}(B^{s_c}) \dots \Phi_{s_1}(B^{s_1})\Phi_1(B)z_t = \Theta_{s_c}(B^{s_c}) \dots \Theta_{s_1}(B^{s_1})\Theta_1(B)a_t . \quad (11)$$

The advantage is using the multiplicative model (11) rather than the general model (9) is, of course, that we are likely to be able to describe the behavior of the series  $\{z_t\}$  in fewer parameters because of the special (seasonal) nature of the series. Furthermore, the special structure allows for much simpler computerized estimation of the model parameters.

### 3. Deterministic Components

The general model (9) is capable of describing many types of stochastic behavior of time series, including changes in level, trends, and pseudo-periodic behavior. If, on the basis of physical considerations, it is known that certain characteristics of a series are deterministic in nature, these components should be explicitly handled as separate terms in the model. For example, if a persistent linear year-to-year trend is present in a series, and it is clear that the trend corresponds, say, to growth of the economy, then this trend should be represented explicitly in the model:

$$\Phi(B)z_t = \alpha t + \Theta(B)a_t$$

where  $\alpha$  is the slope parameter of the trend. Similarly, deterministic trigonometric terms might be warranted:

$$\Phi(B)z_t = \sum_{i=1}^h \sin 2\pi f_i t + \Theta(B)a_t .$$

It is important to note, however, that prior physical considerations, and not a visual examination of the observed data, should be the basis for inclusion of deterministic components. Observations sampled from a process obeying the model (9) can exhibit trends and periodicities that are stochastic in nature (i.e., they may change as the process evolves). To conclude from a visual examination of the data that such trends or periodicities were deterministic would be incorrect.

The use of a multiplicative model for time series suspected of having seasonal behavior is an example of inclusion into the model of another type of deterministic component.

## C. CHARACTERISTICS OF THE TIME SERIES MODEL

This section will describe certain aspects of the behavior of the processes defined by (9), as a function of the parameters ( $\phi$ 's and  $\theta$ 's) of the model:

### 1. Stationarity

A very important class of stochastic processes is the class of *stationary processes*. A *strictly stationary* process is one whose properties are unaffected by a shift in the time origin; i.e., the joint distribution of  $z_t, z_{t-1}, \dots, z_{t-n}$  is the same as the joint distribution of  $z_{t-h}, z_{t-h-1}, \dots, z_{t-h-n}$  for any  $n$  and  $h$ . Roughly speaking, a stationary process is one in which there are no changes in levels or trends, i.e. the process forever fluctuates about a fixed mean. A process is *weakly stationary* of order  $r$  if all its moments of order up to  $r$  are unaffected by a shift in the time origin. The joint probability function of a *Gaussian* process (i.e., process for which the  $a_t$ 's are normally distributed) is characterized (i.e., completely specified) by its first and second moments, and is strictly stationary if it is weakly stationary of order 2.

It can be shown that a process defined by the model (9) is stationary (of order 2) if the zeros of the polynomial  $\Phi(B)$ , considered as a function of the complex variable  $B$ , are outside the unit circle. A stationary process can be represented as

$$z_t = \sum_{i=0}^{\infty} \psi_i a_{t-i}$$

where the function

$$\Psi(B) = \sum_{i=0}^{\infty} \psi_i B^i$$

converges for  $|B| < 1$ .

It is noted that if all the zeros of  $\Theta(B)$  are outside the unit circle, then the process is invertible, i.e., it can be represented as

$$z_t = \sum_{i=1}^{\infty} \pi_i z_{t-i} + a_t ,$$

where the function

$$\Pi(B) = \sum_{i=0}^{\infty} \pi_i B^i$$

converges for  $|B| < 1$ .

## 2. The Autocorrelation Function

Processes that are second-order stationary are very important in time series analysis, and we shall now discuss their second-order properties. The second moments of a process are given by the autocovariance function, defined by

$$\begin{aligned} \gamma_k &= \text{cov}(z_t, z_{t-k}) \\ &= E(z_t - \mu)(z_{t-k} - \mu) \end{aligned}$$

where  $\mu$  denotes the mean of  $z_t$  and E denotes the expectation (expected value) operator. If we normalize the autocovariance function by dividing the variance  $\sigma^2 = \gamma_0$  of the process, we obtain the *autocorrelation* function

$$\rho_k = \gamma_k / \gamma_0 .$$

The probability structure of a stationary Gaussian process is characterized (i.e., completely specified) by knowledge of  $\mu$  and  $\gamma_k$  or, equivalently, by knowledge of  $\mu$ ,  $\sigma^2$ , and  $\rho_k$ . Furthermore, the *model* of the form (9) of a stationary, invertible process is characterized by the autocorrelation function, whether it is Gaussian or not. In fact, if we define the covariance generating function

$$C(B) = \sum_{i=-\infty}^{\infty} \gamma_i B^i$$

and denote  $\sigma_a^2 = \text{var } a_t$ , it can be shown that

$$\gamma_k = \sigma_a^2 \sum_{i=0}^{\infty} \psi_i \psi_{t+k}$$

so that

$$C(B) = \sigma_a^2 L_{\psi}(B) L_{\psi}(B^{-1}) .$$

Thus, a given stationary model (9) has a unique covariance function. Also, it can be proved that there exists only one model of the form (9) *that expresses the current observation in terms of previous history* having a specified covariance function, provided the zeros of  $\Phi(B)$  and  $\Theta(B)$  lie outside the unit circle (i.e., the process is both stationary and invertible). Hence, for stationary invertible processes, the autocorrelation function is of value in indicating the form of the model (9) of the process. If we have a nonstationary process, we must transform it to a stationary process if we wish to use the autocorrelation function as a guide to model selection.

An acceptable estimate of the autocovariance function is the sample autocovariance function, defined by

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

where  $\bar{z}$  denotes the sample mean,

$$\bar{z} = \frac{1}{n} \sum_{t=1}^n z_t$$

and our observed time series is  $z_1, z_2, \dots, z_n$ . An acceptable estimate of the autocorrelation function is the sample autocorrelation function

$$\gamma_k = c_k / c_0.$$

In testing a model, it is desirable to test whether or not the estimated residuals are autocorrelated. Appendix A describes a test of the hypothesis that  $\gamma_k = 0, k=1, 2, \dots, n-1$ , in the case of a normal distribution.

### 3. The Periodogram and the Spectrum

#### a. The Periodogram

A statistic that is useful in detecting deterministic periodicities in data (due to, e.g., sinusoidal terms or seasonality), is the *periodogram*. The periodogram,  $I_n^*(f)$ , is an estimate of the square of the amplitude of the cosine wave of specified frequency,  $f$ , in the time series. It is given by

$$\begin{aligned} I_n^*(f) &= \frac{2}{n} \left| \sum_{t=1}^n (z_t - \bar{z}) e^{-i\pi f t / L} \right|^2 \\ &= 2 \sum_{k=1-n}^{n-1} \left( 1 - \frac{|k|}{n} \right) c_k e^{-i\pi f k / L} \\ &= 2 \left[ c_0 + 2 \sum_{k=1}^{n-1} \left( 1 - \frac{k}{n} \right) c_k \cos \pi f k / L \right] \quad 0 \leq f \leq L \end{aligned}$$

where  $c_i$  is the sample autocovariance function.

Appendix B shows how the estimate of  $I_n^*(f)$  is derived, and provides references for testing the statistical significance of it (i.e., of specified frequencies), in the case of a normal distribution.

#### b. The Spectrum

(Note: Since the spectrum may be somewhat less familiar to the reader than the other quantities defined in this document, it is described here in somewhat greater detail. The spectrum is not as useful a tool for assisting model development as the autocorrelation function.)

If the discrete stochastic process  $\{z_t\}$  of the model (9) is a second order stationary process, it can be represented as

$$z_t = \int_{-L}^L e^{i\pi f t / L} dS(f)$$

where  $S(f)$  is a stochastic process defined over the interval  $(-L, L)$ , where  $L > 0$  is arbitrary. The processes  $z_t$  and  $S(f)$  are different representations of the same model;  $z_t$  corresponds to representing the model in the “time domain,” and  $S(f)$  corresponds to representing the process in the “frequency domain”. While the representations are mathematically equivalent, one or the other is generally more convenient to work with, depending on the nature of the process. The models we consider arise naturally in the time domain and are hence most simply described in it (using the autocorrelation function), but we shall now see an instance in which consideration in the frequency domain (using a function called the spectral density function) is useful.

The right-hand side of the above equation is called the *spectral representation* of the process  $\{z_t\}$ . The second order properties of  $S(f)$  are specified by a function  $P(f) = \text{var } S(f)$ , or equivalently, by the variance ( $\sigma^2$ ) of  $z_t$  and a function  $G(f) = P(f)/\sigma^2$  called the *spectral distribution function*. It can be shown that

$$\gamma_k = \int_{-L}^L e^{i\pi f k / L} dP(f) \quad , \quad k = 0, 1, 2, \dots,$$

or

$$\rho_k = \int_{-L}^L e^{i\pi f k / L} dG(f) \quad .$$

If  $G(f)$  is absolutely continuous, we have

$$\gamma_k = \int_{-L}^L e^{i\pi f k / L} p(f) df,$$

and

$$\rho_k = \int_{-L}^L e^{i\pi f k / L} g(f) df \quad ,$$

where  $p(f)$  is called the spectrum, and  $g(f)$  is called the spectral density function. We see that the autocorrelation function is the Fourier transform of the spectral density function. Further, since  $g(f)$  is an even function, we have

$$\gamma_k = 2 \int_0^L \cos\left(\frac{\pi f k}{L}\right) p(f) df,$$

so that  $\gamma_k$  is, more specifically, twice the finite Fourier cosine transform of  $p(f)$ . We hence have that

$$p(f) = \frac{1}{2L} \left[ \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k \cos\left(\frac{\pi f k}{L}\right) \right], \quad -L \leq f \leq L \quad .$$

In mathematical analysis, it is customary to choose  $L = \pi$  so that the frequency  $f$  is measured in units of radians per unit time. Because of the natural occurrence of the time dimension in our models, however, it is more natural to choose  $L = 1/2$ , so that  $f$  is in cycles per unit time. In this case, the period corresponding to frequency  $f$  is simply the reciprocal of  $f$ .

Note that

$$\gamma_0 = \sigma^2 = \int_{-\frac{1}{2}}^{\frac{1}{2}} p(f) df$$

or

$$\rho_0 = 1 = \int_{-\frac{1}{2}}^{\frac{1}{2}} g(f) df .$$

Since  $g(f)$  is positive, it has the properties of a probability density function. The spectral density function in fact indicates the distribution of the variance of the time series over the frequency range.

An alternative definition of the spectrum is

$$\rho^*(f) = 2 \left[ \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k \cos\left(\frac{\pi f k}{L}\right) \right] , \quad 0 \leq f \leq L$$

where

$$\gamma_k = \frac{1}{2L} \int_0^L p^*(f) \cos\left(\frac{\pi f k}{L}\right) df , \quad k = 0, 1, 2, \dots,$$

and  $L$  is taken equal to  $\frac{1}{2}$ . That is, the coefficient of the term in brackets in the definition of the power spectrum is 2 rather than  $1/(2L)$ . this definition (with  $L = \frac{1}{2}$ ) has the property that

$$\rho_0 = \sigma^2 = \int_0^{\frac{1}{2}} p^*(f) df ,$$

so that the spectral density function is

$$g(f) = p^*(f) / \sigma^2 .$$

This definition arises naturally since

$$\lim_{n \rightarrow \infty} E[I_n^*(f)] = p^*(f)$$

that is, the power spectrum is the limit of the expected value of the periodogram as the sample size increases.

For a white noise sequence,  $\gamma_i = 0$  for  $i = 1, 2, \dots$ , and so its spectrum is

$$p(f) = \frac{\gamma_0}{2L} = \frac{\sigma^2}{2L} , \quad L \leq f \leq L ,$$

i.e., the spectrum is “flat” (equal to a constant). For a sampled white noise sequence, the estimate of the spectrum (given in Appendix C) will, of course, have some variability. Appendix C describes a test of the spectrum for departures from whiteness, assuming a normal distribution. This test complements the test for zero autocorrelation using the autocorrelation function.

It is important to recognize the distinction between the periodogram and an estimate of the spectrum. The periodogram is intended to indicate the presence of *deterministic* periodic behavior, such as an annually recurring phenomenon, whereas an estimate of the spectrum is intended to indicate the presence of *stochastic* periodic behavior, such as an irregular cycle of some sort. Whereas the periodogram is intended to measure the significance of exactly specified frequencies, the spectral estimate attempts to measure the distribution of frequencies over a specified range. This difference in the intended usage of the two statistics in fact results in the inclusion of certain smoothing constants in the spectral estimate which are not present in the periodogram.

#### 4. Behavior of a Process as a Function of its Model Parameters

The behavior of a process defined by the model (9) can be described in general terms as a function of the parameters of the model. If the zeros of the polynomial  $\Phi(B)$ , considered as a function of the complex variable  $B$ , are inside the unit circle, the process exhibits explosive or increasing

oscillating behavior. If the zeros are outside the unit circle, then the process is stationary. If the roots are on the unit circle, then the process is nonstationary, but exhibits a homogeneous behavior as it evolves. Roughly speaking, the local behavior of the series in this case is independent of the level of the process. In particular, trends and changes in level are possible in this case, and so this type of nonstationary model is potentially useful for modeling many types of naturally occurring processes, such as sales and prices.

In particular, we shall restrict consideration to processes of the following form:

$$\Phi(B)\nabla_{s_c}^{d_{s_c}} \dots \nabla_{s_1}^{d_{s_1}} \nabla^{d_{z_t}} = \Theta(B)a_t \quad (12)$$

where the operators  $\Phi(B)$  and  $\Theta(B)$  are assumed to have all their zeros outside the unit circle. (The invertibility restriction on  $\Theta(B)$  is not really a restriction at all. If a noninvertible representation of the process exists, then so does an invertible representation (in terms of a different white noise sequence).)

All of the zeros of the autoregressive operator lying on the unit circle are contained in the difference operators (e.g.,  $\nabla = 1 - B$  has the zero  $B = 1$ ). The variate

$$w_t = \nabla_{s_c}^{d_{s_c}} \dots \nabla_{s_1}^{d_{s_1}} \nabla^{d_{z_t}} z_t$$

is thus a stationary variate, and its model  $\Phi(B)w_t = \Theta(B)a_t$  is hence characterized by its autocovariance function or spectrum. The autocorrelation function or spectral density function of  $w_t$  can hence be used as a guide to the form of the  $\Phi$  and  $\Theta$  polynomials.

The form of the polynomial  $\Phi(B)$  has a significant effect on the behavior of the process. If the variance of the  $a_t$ 's is small, then the process will behave much like the difference equation

$$\Phi(B)z_t^* = 0.$$

The solution of this difference equation will indicate the general behavior of the process. In fact, for a Gaussian process, this solution is the mean. If the zeros of  $\Phi(B)$  are complex, for example, then the solution  $z_t^*$  involves sines and cosines, and so  $z_t$  will exhibit adaptive cyclic behavior. Appendix D provides some solutions to simple linear difference equations.

## ***D. STATISTICAL ANALYSIS OF THE TIME SERIES MODEL***

### **1. Tentative Model Selection**

#### **a. Differencing for Stationarity**

As indicated earlier, one of the first steps in model building is to decide upon a tentative subclass of models to be analyzed. In particular, we wish to determine the nature of the  $\Phi$  and  $\Theta$  polynomials: the number of seasonal components, and the number of terms in the component polynomials. The theory described above indicates that the autocorrelation function and the spectrum should be of some help if we are working with a stationary process. With this in mind, we hence seek to transform the variate  $z_t$ , which may be nonstationary, to a stationary variate. Nonstationarity is indicated by an autocorrelation function that does not die out. Nonstationarity is allowed for in the model (12) in the form of differences, and so we proceed to take differences until we obtain a variate

$$w_t = \nabla_{s_c}^{d_{s_c}} \dots \nabla_{s_c}^{d_{s_c}} \nabla^{d_{z_t}} z_t$$

whose autocorrelation function dies out.

## b. The autocorrelation Function and Partial Autocorrelation Function

### (1) Pure Moving Average Process

It can be proved that the autocorrelation function of a pure moving average process

$$w_t = \Theta(B)a_t$$

cuts off at the order of the process. Thus if

$$\Theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q$$

then  $\gamma_q$  is nonzero, but  $\gamma_i = 0$  for  $i > q$ .

Hence, if the sample autocorrelations  $\hat{\gamma}_i$  are not statistically significantly different from zero after some lag  $q'$ , we can tentatively entertain a pure moving average process of order  $q'$  as our model in the variate  $w_t$ .

### (2) Pure Autoregressive Process

Let us define the partial autocorrelation coefficient of lag  $h$  to be the last coefficient,  $\pi_h$ , of a pure autoregressive process of order  $h$ . Further, let us define the  $h$ -th sample partial autocorrelation coefficient  $p_h$  as the estimate of the last ( $h$ -th) coefficient in the autoregressive model of order  $h$  fitted (by the method of least squares) to the time series. (The least-squares estimate will be defined in a later section). Then, by definition, if we have a pure autoregressive process, the partial autocorrelation coefficient cuts off at the order of the process. Thus if

$$\Phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$$

then  $\pi_p$  is nonzero, but  $\pi_i = 0$  for  $i > p$ . Hence if the sample partial autocorrelation coefficients are not statistically significantly different from zero after some lag  $p'$  we can tentatively entertain a pure autoregressive process of order  $p'$  as our model in the variate  $w_t$ .

### (3) Mixed Autoregressive Moving Average Model

If both  $\phi^i$ 's and  $\theta^j$ 's are present in the model (in  $w_t$ ), then neither the autocorrelation nor partial autocorrelation coefficients cuts off; rather, they both die out. The behavior of the functions is still a useful guide to the form of the model, however. Box and Jenkins provide a table illustrating this behavior as a function of  $p$  and  $q$ .

## c. The Periodogram and the Spectrum

The periodogram is used to detect the presence of deterministic components in the process. As noted earlier, caution should be taken so as not to remove stochastic periodicities by relating tentatively identified deterministic components to physical phenomena.

The spectrum, although mathematically equivalent to the autocorrelation function, is not nearly as helpful in guiding the selection of parametric models. The principal use of the spectrum will be as a test for whiteness of residuals, to be described later.

## 2. Estimation

### a. Least-Squares Estimates

The procedure used to derive estimates of the parameters of the model is that of least-squares. The least-squares estimates are used because they have a number of desirable statistical properties, and they are easily computed. Provided the residuals are normally distributed, uncorrelated, and have constant variance, they are asymptotically (i.e., for large sample sizes) efficient (i.e., their sampling variance is less than the sampling variance of any other linear estimates) and consistent (for large samples there is, loosely speaking, a high probability that the parameter estimate is close to the parameter being estimated).

### b. Maximum Likelihood Estimates

An alternative method to least-squares for obtaining estimates is the method of maximum likelihood. In this procedure, a distribution is assumed for the  $a_i$ 's. This distribution involves the unknown parameters of the model. Those values are then chosen for the parameters that maximize the probability ("likelihood") of having observed the particular sample of  $a_i$ 's that were actually observed. Although maximum likelihood estimates are identical to least-squares estimates in the case of normally distributed  $a_i$ 's there are a number of advantages associated with knowledge of the entire likelihood function (as a function of the parameters):

1. Restriction of the parameters to certain regions (of stationarity or invertibility) are easily handled.
2. Alternative models that, statistically speaking, are about as good as the least-squares model, may be identified.
3. Parameter redundancies, such as a common factor in  $\Phi(B)$  and  $\Theta(B)$ , would be recognized (the maximum would tend to lie along a line or plane).

## 3. Diagnostic Checking

### a. Hypothesis Testing

In the course of analysis of a tentatively proposed model, it is necessary to make various statistical tests of hypotheses concerning the values of parameters or the shapes of distributions. Such tests of hypotheses assume a particular distribution for the  $a_i$ 's. Note that none of the estimates (except the maximum likelihood estimates) depends on the distribution form. In what follows, we shall present tests based on the assumption of a normal distribution of the  $a_i$ 's. Usually, the tests will still be appropriate if the  $a_i$ 's deviate somewhat from normality. In any event, it is possible to test the hypothesis of normality of the residuals ( $a_i$ 's) by means of a Kolmogorov-Smirnov test of goodness of fit.

## b. Significance of Parameters

In model building, it is generally less undesirable to risk introduction of an unnecessary parameter into the model, than to omit a necessary one. Nevertheless, it is desirable to be able to test whether or not the estimated  $\phi$ 's and  $\theta$ 's of the fitted model

$$\hat{\Phi}(B)w_t = \hat{\Theta}(B)a_t$$

are statistically significantly different from zero.

Let us denote the vector of all estimates by  $\underline{\beta}$ , i.e.,

$$\underline{\beta}' = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q).$$

The covariance matrix,  $S$ , of  $\underline{\beta}$  is computed as described in Appendix E. Let us define Hotelling's  $T^2$  statistic:

$$T^2 = (\underline{\beta} - \underline{\beta}_0)' S^{-1} (\underline{\beta} - \underline{\beta}_0)$$

where  $\underline{\beta}_0$  is the test value of  $\underline{\beta}$ , in this case, the vector of  $p+q$  zeros. Then, under the hypothesis  $\underline{\beta} = \underline{\beta}_0$ , the quantity

$$F = \frac{T^2}{p+q} \cdot \frac{n-p-q}{n-1}$$

has an  $F$ -distribution with  $p+q$  and  $n-p-q$  degrees of freedom, where  $n$  is the number of observations. To perform a 95% significance level test of the hypothesis

$$\underline{\beta} = \underline{0},$$

we simply compute  $F$  and reject the hypothesis if

$$F > F_{.95}(p+q, n-p-q),$$

where  $F_{.95}(p+q, n-p-q)$  is the 95% critical value of the  $F$ -distribution with parameters  $p+q$  and  $n-p-q$ .

## c. Analysis of Residuals

### (1) Autocorrelation Analysis

Perhaps the most important test of the statistical adequacy of a fitted model

$$\hat{\Phi}(B)w_t = \hat{\Theta}(B)a_t$$

is a test of whether or not the model residuals  $\hat{a}_t$  corresponding to the estimates  $\hat{\Phi}(B)$  and  $\hat{\Theta}(B)$  are white noise. (The model residuals  $\hat{a}_t = \hat{\Phi}^{-1}(B)\hat{\Theta}(B)w_t$  are estimates of the (unobservable)  $a_t$ 's.)

Although even for a "correct" model these estimates are slightly correlated, for "large" samples their (sample) autocorrelation function approaches the autocorrelation function of the  $a_t$ 's.) Evidence to the contrary indicates that the model is incorrect or incomplete, and should be revised. Appendix A presents a test for whiteness in terms of the autocorrelation function.

The autocorrelation behavior of the residuals of a fitted model can be used as a guide to revising the model. For example, suppose that we have tentatively fitted the model

$$\hat{\Phi}(B)w_t = \hat{\Theta}(B)e_t$$

and that the model residuals  $\hat{e}_t$  are correlated. If the correlation function of the  $\hat{e}_t$ 's suggests a model of the form

$$\hat{\Phi}^*(B)e_t = \hat{\Theta}^*(B)a_t$$

then the new model

$$\Phi(B)\Phi^*(B)w_t = \Theta(B)\Theta^*(B)a_t$$

should be examined.

## (2) *Periodogram Analysis*

If seasonal components are suspected, computation of the periodogram of the residuals of a fitted model should indicate whether or not the model should be revised to reflect such behavior. See Appendix B.

## (3) *Spectral Analysis*

Appendix C presents a test for whiteness of the residuals of a model in terms of the spectrum.

### **d. Over-fitting**

After we have derived a model whose residuals satisfy the preceding tests, we should allow for additional parameters in the fitted model, and determine whether or not their estimates are statistically significantly different from zero. If they are, then there is cause for concern that we have not identified the model correctly. For example, if we have fitted the model

$$(1 - \phi_1 B - \phi_2 B^2)w_t = a_t,$$

we should also examine

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3)w_t = a_t$$

or

$$(1 - \phi_1 B - \phi_2 B^2)w_t = a_t - \theta_1 a_{t-1}$$

to see whether or not the additional parameters are significant.

### **e. Independent Estimates from Different Data**

It is possible that the parameters of a process being modeled are not constant over time. If so, we are in danger of constructing a model that describes a certain period of history well, but that is of little predictive value. For this reason, it is desirable to derive parameter estimates from separate sections of time series history, and to compare the results.

In actual practice, it would be best to review parameter estimates as new data become available, to guard against model changes. (If the parameters must be changed frequently, this should be taken as evidence that the model is not adequate to describe the process well. In such a case, a different time series representation should be sought, or perhaps an “adaptive” model, such as a Kalman filter.)

## E. SIMULATION AND FORECASTING

### 1. Simulation

In some instances, a model is constructed in order to allow simulation of a process. For example, it could be desired to examine the effect of sales variability on a new inventory policy, starting from the current sales position. In such a case, we not only need to know the parameters of the model

$$\Phi(B)\nabla_{s_c}^{d_{s_c}} \dots \nabla_{s_1}^{d_{s_1}} \nabla^d z_t = \Theta(B)a_t$$

but also the variance of  $a_t$  and the distribution of  $a_t$ .

The variance of  $a_t$  is estimated as part of the least-squares estimation procedure. To estimate the form of the distribution of  $a_t$ , the model residuals ( $a_t$ 's) corresponding to the estimated parameters should be used to construct a histogram. The shape of the histogram will generally suggest a parametric class of distributions (such as normal) from which the  $a_t$ 's could be regarded as being sampled. The Kolmogorov-Smirnov test of goodness of fit could be used to test an hypothesis regarding a choice for the distribution of the  $a_t$ 's.

To simulate the process

$$(1 - \phi_1 B - \dots - \phi_p B^p) z_t = (1 - \theta_1 B - \dots - \theta_q B^q) a_t$$

we simply write it in the form

$$z_t = \phi_1 z_{t-1} + \dots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q} .$$

If we have past history,  $z_{t-1}, \dots, z_{t-p}$ , from an actual realization of the time series, then we substitute these values in the above expression. The value  $a_t$  is drawn from the identified distribution of the residuals. The values  $a_{t-1}, a_{t-2}, \dots, a_{t-q}$ , must be estimated from the past data. A simple recursive procedure for doing so will be described later in the section on forecasting. Thus all of the quantities on the right-hand side of the above expression are known, and we can compute the value for  $z_t$ . The new values  $a_t$  and  $z_t$  are in turn used, with a new sampled value  $a_{t+1}$ , to compute  $z_{t+1}$ , and so on.

If we have no past history, then it is necessary to determine initial values for  $z_1, \dots, z_p$  and  $a_1, \dots, a_q$ . The initial  $a_t$ 's are obtained simply by sampling. The initial  $z_t$ 's should be drawn from a distribution of initial values. Such a distribution would be determined by analysis of the initial values. Such a distribution would be determined by analysis of the initial values of actual time series, or by subjective considerations. Note, of course, that the initial  $z_t$ 's are correlated, so that they cannot be independently sampled. If the  $a_t$ 's are normally distributed, then a regression model for  $z_2$  conditional on  $z_1$ , and  $z_3$  conditional on  $z_2$  and  $z_1$ , is a legitimate procedure for sampling the initial values  $z_1, \dots, z_p$ .

### 2. Forecasting

A useful property of the time series model (9) is that it is easy to derive a forecasting function for it. The forecasting function can be expressed simply; it has the property that it minimizes the expected squared deviation of the forecast from the actual value.

## a. Derivation of the Least-Squares Forecaster

(Note: The reader who is not interested in the details of the derivation of the least-squares forecaster may omit reading this section.)

Solving the difference equation represented by the model (9), we may write

$$\begin{aligned} z_t &= C(t) + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots \\ &= C(t) + L_\psi(B)a_t \end{aligned}$$

where  $C(t)$  is the complementary function (general solution of the homogeneous equation), and  $L_\psi(B)a_t$  is a particular solution of the complete equation. Consider a forecaster  $z_t^*(\ell)$  of  $z_{t+\ell}$  that is a linear function of the  $a_i$ 's:

$$\begin{aligned} z_t^*(\ell) &= C(t+\ell) + \psi_0^* a_t + \psi_1^* a_{t-1} + \psi_2^* a_{t-2} + \dots \\ &= C(t+\ell) + L_{\psi^*}(B)a_t \end{aligned}$$

Since

$$z_{t+\ell} = C(t+\ell) + a_{t+\ell} + \psi_1 a_{t+\ell-1} + \dots + \psi_{\ell-1} a_{t+1} + \psi_\ell a_t + \psi_{\ell+1} a_{t-1} + \dots,$$

the forecast error is

$$\begin{aligned} e_t(\ell) &= z_{t+\ell} - z_t^*(\ell) \\ &= a_{t+\ell} + \psi_1 a_{t+\ell-1} + \dots + \psi_{\ell-1} a_{t+1} + (\psi_\ell - \psi_0^*) a_t + (\psi_{\ell+1} - \psi_1^*) a_{t-1} + \dots \end{aligned}$$

Assuming that the  $a_i$ 's are uncorrelated with mean zero and variance  $\sigma^2$  the mean squared error is

$$\begin{aligned} E(e_t(\ell)) &= E(z_{t+\ell} - z_t^*(\ell))^2 \\ &= (1 + \psi_1^2 + \dots + \psi_{\ell-1}^2) \sigma^2 + \sum_{i=0}^{\infty} (\psi_{\ell+i} - \psi_i^*)^2 \sigma^2 \end{aligned}$$

This quantity is minimized by setting  $\psi_i^* = \psi_{\ell+i}$ . Hence the least-squares forecaster is

$$\hat{z}_t(\ell) = C(t+\ell) + \psi_\ell a_t + \psi_{\ell+1} a_{t-1} + \psi_{\ell+2} a_{t-2} + \dots$$

Since

$$\Phi(B)z_t = \Theta(B)a_t$$

and

$$\Phi(B)C(t) = 0$$

and

$$z_t = C(t) + L_\psi(B)a_t,$$

we have

$$\Phi(B)z_t = \Phi(B)L_\psi(B)a_t,$$

so that

$$\Phi(B)L_\psi(B) = \Theta(B)$$

or

$$(1 - \phi_1 B - \dots - \phi_p B^p)(1 + \psi_1 + \psi_2 B + \dots) = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q) \ .$$

Identifying coefficients of powers of  $B$ , and denoting  $\psi_0 = 1$ , we have

$$\psi_1 = \phi_1 - \theta_1$$

$$\psi_2 = \phi_1 \psi_1 + \phi_2 - \theta_2$$

...

$$\psi_q = \phi_1 \psi_{q-1} + \phi_2 \psi_{q-2} + \dots + \phi_p \psi_{q-p} - \theta_q \quad (p \leq q)$$

(or

$$\psi_p = \phi_1 \psi_{p-1} + \phi_2 \psi_{p-2} + \dots + \phi_p \psi_0 \quad \text{if } p > q)$$

$$\psi_r = \phi_1 \psi_{r-1} + \phi_2 \psi_{r-2} + \dots + \phi_p \psi_{r-p} \quad (r > \max(p, q)).$$

Thus the  $\psi_r$ 's may be calculated recursively.

The variance of the error in the forecast  $\hat{z}_t(\ell)$  is given by

$$\text{var } e_t(\ell) = (1 + \psi_1^2 + \dots + \psi_{t-1}^2) \sigma^2.$$

Thus the variance of the one-ahead forecast is simply  $\sigma^2$ . The variance of forecasts further out approaches  $\text{var}(z_t)$ .

## b. Computation of the Forecast

Note that

$$z_{t+1} - \hat{z}_t(1) = a_{t+1} \quad (13)$$

and that

$$E(z_{t+k}) = \hat{z}_t(k)$$

where  $E(z_{t+k})$  denotes the conditional expectation of  $z_{t+k}$ , given  $z_t, z_{t-1}, \dots$

In actual practice, the  $\psi$ 's are not necessary to compute the forecast  $\hat{z}_t(\ell)$ . All that is required is to write down the model (9) with all terms but  $\hat{z}_{t+\ell}$  on the right-hand side. We then substitute in the right-hand side, as required, the following:

1.  $z_t, z_{t-1}, \dots$ , (observed);
2.  $a_t, a_{t-1}, \dots$ , (computed from (13));
3.  $\hat{z}_t(k)$  for  $z_{t+k}$ ,  $1 \leq k < \ell$ ;
4. zeros for  $a_{t+k}$ ,  $1 \leq k \leq \ell$ .

Thus we simply substitute known quantities and replace unknown quantities by their expected values.

It is noted that

$$\hat{z}_{t+1}(\ell) = z_t(\ell + 1) + \psi_\ell a_{t+1},$$

so that, once  $z_{t+1}$  is known, the forecasts  $\hat{z}_{t+1}(\ell)$  can be readily computed from the quantities

$$\hat{z}_t(\ell + 1) \text{ and } a_{t+1} = z_{t+1} - \hat{z}_t(1).$$

Recall that in the section on simulation we needed estimates of the past  $a_t$ 's in order to start the simulation process. The formula (13) provides the means for obtaining these estimates. That is, we simply start forecasting one step ahead from the beginning of our historical time series (assuming zeros for  $a_1, a_0, a_{-1}, a_{-2}, \dots$ ). For each time  $t$  the difference between the realized value of the series at time  $t$  and the forecast  $\hat{z}_{t-1}(1)_t$  of  $z_t$  made at time  $t-1$  is an estimate of  $a_t$ .

### c. Eventual Forecast Function

After  $\ell > q$  steps into the future, the forecast  $\hat{z}_t(\ell)$  is a solution of the autoregressive part of the process set equal to zero:

$$\Phi(B)z_t = 0 \quad . \quad (14)$$

This solution is called the *eventual forecast function*. Note that if  $q > (p+d) = P$ , say, then the eventual forecast function is the unique curve (whose form is specified by the solution to (14) passing through the last  $P$  of the forecasts  $\hat{z}_t(q), \dots, \hat{z}_t(q - P + 1)$ . If  $q \leq P$ , then both the actual forecast and the eventual forecast function are the unique curve (determined by  $\Phi(B)$ ) that passes through the first  $P$  forecasts  $\hat{z}_t(1), \dots, \hat{z}_t(P)$ .

For example, for autoregressive functions of the form

$$\Phi(B) = (1 - B)^2 \quad ,$$

the eventual forecast function is given by

$$\hat{z}_t(\ell) = b_{1t} + b_{2t}\ell \quad ,$$

i.e., a straight line. For

$$\Phi(B) = (1 - cB)^2$$

we have

$$\hat{z}_t(\ell) = (b_{1t} + b_{2t}\ell)c^\ell \quad .$$

## II. THE TIME SERIES ANALYSIS PROGRAM *TIMES*

The time series analysis program *TIMES* is capable of performing all the computations described in the preceding sections, except the computation of the likelihood function. The *TIMES* program consists of two subprograms: *ESTIMATE*, which estimates the model parameters and analyzes the model residuals; and *PROJECTOR*, which computes projections. The *PROJECTOR* program may be used in either of two modes: in the *SIMULATE* mode, to compute simulations; and in the *FORECAST* mode, to compute forecasts. We shall now describe the capabilities of these programs in somewhat more detail.

### A. THE ESTIMATION PROGRAM “ESTIMATE”

The *ESTIMATE* program estimates the parameters of a mixed autoregressive moving average time series model having multiplicative seasonal components. Sine terms, cosine terms, and exogenous variates may also be included in the models. In addition to estimating parameters, the program also computes the mean, variance, histogram, autocorrelation function, partial autocorrelation function, and spectrum of the residuals of the estimated model.

The program is designed to work with the actual data, or to transform to logarithms. To run the program, the user must specify the number,  $c$ , of multiplicative (seasonal) components additional to the basic component, the periods  $s_c, \dots, s_1$  of the seasons, the respective difference parameters,  $d_{s_c}, \dots, d_{s_1}, d$ , the orders of the  $\Phi$  polynomials  $p_{s_c}, \dots, p_{s_1}, p$ , and the orders of the  $\Theta$  polynomials,  $q_{s_c}, \dots, q_{s_1}, q$ .

In addition, it is possible to use the operators

$$(1 - \zeta_{s_c} B^{s_c}), \dots, (1 - \zeta_{s_1} B^{s_1}), (1 - \zeta B)$$

instead of

$$\nabla_{s_c}, \dots, \nabla_{s_1}, \nabla_{s_2}, \dots$$

in which case the  $\zeta^i s$  must be specified as different from unities. It is possible to combine several different time series (of equal length) as the basis for estimating the overall model parameters, and an option is available for subtracting the mean of each series from the observations of the series. Further, it is possible to include deterministic sine and cosine terms, in which case the frequencies must be specified. Finally, one can include exogenous variables, whose values are the same or different for corresponding observations of different series, and an overall mean. In these latter cases, the program estimates the mean, the coefficients of the sine and cosine terms, and the coefficients of the exogenous variates. Thus the model analyzed by *ESTIMATE* is

$$\begin{aligned} & \Phi_{s_c}(B^{s_c}) \dots \Phi_{s_1}(B^{s_1}) \Phi(B) \nabla_{s_c}^{d_{s_c}} \dots \nabla_{s_1}^{d_{s_1}} \nabla^d z_{ij} \\ & = \Theta_{s_c}(B^{s_c}) \dots \Theta_{s_1}(B^{s_1}) \Theta(B) a_{ij} + \mu \\ & + w_1 \sin 2\pi F_1 t + \dots + w_h \sin 2\pi F_h t \\ & + v_1 \cos 2\pi G_1 t + \dots + v_k \cos 2\pi G_k t \\ & + t_1 X_{1t} + \dots + t_f X_{ft} \\ & + u_1 Y_{1t} + \dots + u_g Y_{gt} \quad t = 1, \dots, m; j = 1, \dots, n \end{aligned}$$

where  $z_{ij}$  denotes the  $t$ -th observation in the  $j$ -th series, and where the user specifies the  $s$ 's,  $d$ 's,  $F$ 's ( $0 < F_i \leq 1/2$ ),  $G$ 's ( $0 < G_i \leq 1/2$ ), and the  $X$ 's and  $Y$ 's. As noted earlier, the operators  $\nabla_{s_i} = 1 - B^{s_i}$  can be replaced by  $1 - \zeta_{s_i} B^{s_i}$  by specifying the  $\zeta^i s$  different from unity.

In the case of a nonlinear model ( $\theta$ 's present, or more than one  $\Phi$  polynomial), it is necessary to specify an iteration cutoff value (e.g., .005), a maximum number of iterations (e.g., 20), and a stepsize fraction (e.g., .5).

Specifically, the program performs the following computations:

1. Estimation of parameters.
2. Covariance matrix of parameter estimates.
3. Hotelling's  $T^2$  test of significance of all parameter estimates.
4. Calculation of residuals; histogram of residuals.
5. Mean and variance of residuals.
6. Autocorrelation and partial autocorrelation functions of residuals; correlogram (plot of autocorrelation function).
7. Chi-square test of significance of autocorrelations.
8. Periodogram and estimate of spectrum of residuals.
9. Fisher's test of significance of periodogram.
10. Kolmogoroff-Smirnov test of significance of spectrum.

Based on an examination of the above statistics, the user decides whether or not the model is an adequate statistical representation of the stochastic process generating the time series data. If it is not, then he modifies its structure in accordance with the theory outlined in the text, and runs the program again.

A separate manual (Reference 2) describes the procedures for using the *ESTIMATE* program.

## **B. THE PROJECTION PROGRAM “PROJECTOR”**

Once a statistically adequate model had been derived, it can be used to simulate future realizations, or outcomes, of the stochastic process being modeled. The *PROJECTOR* program computes these simulations. Reference 2 describes the procedures for using the *PROJECTOR* program. The *PROJECTOR* program can simulate any model of the form specified above.

The estimated time series model can also be used to compute forecasts, or predictions, of future observations of the time series. The program *PROJECTOR* computes these forecasts. Also, upper and lower one-standard deviation limits for each forecast value are computed by the program. The user inputs the parameter estimates computed by *ESTIMATE*. The *PROJECTOR* program can forecast any model of the form given above. Reference 2 describes the procedures for using the *PROJECTOR* program.

## **REFERENCES**

1. G.E.P. Box and G.M. Jenkins, *Time Series Analysis, Forecasting, and Control*, Holden-Day, Inc., San Francisco, Calif., 1970, 1976, 1994 (with Gregory Reinsel).
2. *TIMES Reference Manual, Volume II, User's Manual*, Lambda Corporation, Arlington, Virginia, Revised March, 1971.

## APPENDIX A. AUTOCORRELATION ANALYSIS

The usefulness of the autocorrelation function as a guide in model selection has already been described in the text. This appendix will present tests of the significance of the individual autocorrelations and of the autocorrelation function as a whole. The latter test is, of course, a test of “whiteness”.

The autocovariance function,  $\gamma_k$ , of a discrete stochastic process  $\{a_t\}$  is defined as

$$\gamma_k = \text{cov}(a_t, a_{t+k})$$

i.e., it is the covariance between two values,  $a_t$  and  $a_{t+k}$  considered as a function of their distance apart,  $k$ . The time difference  $k$  is called the lag.

We have

$$\gamma_0 = \text{cov}(a_t, a_t) = \sigma^2$$

where  $\sigma^2 = \text{var } a_t$ . The autocorrelation function  $\rho_k$  is defined as

$$\rho_k = \gamma_k / \gamma_0 = \gamma_k / \sigma^2.$$

(A normal stationary process is completely specified by its mean and its autocovariance function.) An unbiased estimate of the autocorrelation function,  $\rho_k$ , is the sample autocorrelation function

$$r_k = c_k / c_0$$

where

$$c_k = \frac{1}{n-k} \sum_{t=1}^{n-k} (a_t - \bar{a})(a_{t+k} - \bar{a}),$$

and

$$\bar{a} = \frac{1}{n} \sum_{t=1}^n a_t.$$

If  $r_k$  denotes the  $k$ -th correlation coefficient, then Bartlett (Reference A-1) has proved that

$$\text{cov}(r_k, r_{k+1}) \approx \frac{1}{n_k} \sum_{i=-\infty}^{\infty} \rho_i \rho_{i+k}$$

and

$$\text{var}(r_k) \approx \frac{1}{n_k} \left( 1 + 2 \sum_{i=1}^{\infty} \rho_i^2 \right),$$

where  $n_k$  is the size of the sample used to estimate  $r_k$ .

Under the white noise hypothesis, we have

$$\text{cov}(r_k, r_{k+1}) \approx 0$$

$$\text{var}(r_k) \approx \frac{1}{n_k}.$$

The sample autocorrelations are thus approximately a sample of uncorrelated random variables with mean 0, variance  $1/n_k$ . (Since the  $a_t$ 's are computed from estimated values for the  $\hat{\phi}$ 's and  $\hat{\theta}$ 's, this approximation is not very good for autocorrelations of small lag. The quantity  $1/n_k$  is an approximate upper bound, however. See Reference A-3 for discussion.) For samples of 30 or more, they are approximately normally distributed (Reference A-2). Whenever the residuals are based on estimated parameters ( $\hat{\phi}$ 's and  $\hat{\theta}$ 's) rather than true values, biases are introduced into the distribution of the autocorrelations. These biases are removed by replacing  $n_k$  by  $n_k - (\text{number of estimated parameters})$ .

Using the preceding results, a test statistic for the hypothesis that the process is a pure moving average process of order  $k-1$  is given by

$$s_k = \frac{r_k}{\left( \frac{1}{n_k} \left( 1 + 2 \sum_{i=1}^{k-1} r_i^2 \right) \right)^{1/2}}$$

where  $s_k$  is approximately distributed as a unit normal variate. Hence  $s_k \leq 1.96$  would be a 95% acceptance region. In order to test for whiteness, we test the significance of  $p$  autocorrelation coefficients using the statistic

$$x^2 = \sum_{k=1}^p n_k r_k^2$$

where  $x^2$  is distributed as a  $\chi^2$  variate with  $p-v$  degrees of freedom, and  $v$  is the number of estimated parameters ( $\hat{\phi}$ 's and  $\hat{\theta}$ 's).

The hypothesis of zero autocorrelations is rejected for  $x^2 > \chi_{crit}^2$ . It is helpful to plot the set of autocorrelations on a graph, showing upper and lower significance levels for the individual autocorrelations. Such a plot is called a *correlogram*.

## REFERENCES

- A-1. Bartlett, M.S., "On the Theoretical Specification and Sampling Properties of Autocorrelated Time Series," *Journal of the Royal Statistical Society, Series B*, Volume 8, 1946.
- A-2. Anderson, R.L., "Distribution of the Serial Correlation Coefficient," *Annals of Mathematical Statistics*, Volume 13, 1942.
- A-3. Box, G.E.P. and Pierce, D.A., "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models," *Journal of the American Statistical Association*, Volume 65, No. 332, December, 1970

## APPENDIX B. PERIODOGRAM ANALYSIS

This Appendix will describe the periodogram, a statistic designed to detect the presence of deterministic periodicities in a time series.

Suppose that, instead of the model (9), the appropriate time series model for sales is of the form

$$z_t = \phi_1 z_{t-1} + \dots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q} \\ + \alpha_o + \sum_{i=1}^r (\alpha_i \cos 2\pi f_i t + \beta_i \sin 2\pi f_i t) \quad (\text{B-1})$$

i.e., deterministic components of periods  $1/f_i$ ,  $i = 1, \dots, r$  are present. This section will show how it is possible to identify these frequencies using the raw periodogram (see Box and Jenkins [Reference B-1] or Wilks [Reference B-2] for further discussion). For simplicity of discussion, we shall ignore the  $\phi$  and  $\theta$  terms and write the model (B-1) as

$$z_t = \alpha_o + \sum_{i=1}^r (\alpha_i \cos 2\pi f_i t + \beta_i \sin 2\pi f_i t) t. \quad (\text{B-2})$$

In actuality, all parameters must be estimated simultaneously.

Let us suppose for convenience that  $n = 2r+1$ , and that  $n$  corresponds to a complete number of cycles of each frequency  $f_i$  so that we may set

$$f_i = i/n, \quad i = 1, 2, \dots, r.$$

Note that we are trying to estimate those harmonics ( $f_i$ ) of the fundamental frequency  $1/(2r+1)$  that correspond to periods greater than twice the time interval length.

Denoting  $c_{it} = \cos 2\pi f_i t$  and  $s_{it} = \sin 2\pi f_i t$ , the model (B-2) becomes

$$z_t = \alpha_o + \sum_{i=1}^r \alpha_i c_{it} + \sum_{i=1}^r \beta_i s_{it} + a_t. \quad (\text{B-3})$$

The least-squares estimates of  $\alpha_o$ ,  $\alpha_i$ , and  $\beta_i$  can be shown to be

$$a_o = \frac{1}{n} \sum_{t=1}^n z_t \\ a_i = \frac{2}{n} \sum_{t=1}^n z_t c_{it} \\ b_i = \frac{2}{n} \sum_{t=1}^n z_t s_{it}.$$

The model (B-3) may be written

$$z_t = \alpha_o + \sum_{i=1}^r \sqrt{\alpha_i^2 + \beta_i^2} \cos(f_i t + \phi_i), \quad (\text{B-4})$$

where  $\tan \phi_i = -\beta_i / \alpha_i$ . The quantity

$$A(f_i) = (\alpha_i^2 + \beta_i^2)$$

is an estimate of the square of the amplitude of the cosine wave of frequency  $f_i$  in  $z_t$ . Defining  $j = \sqrt{-1}$ , we may write

$$A(f_i) = (a_i - jb_i)(a_i + jb_i) .$$

Since

$$\begin{aligned} a_i - jb_i &= \frac{2}{n} \sum_{t=1}^n z_t (c_{it} - js_{it}) \\ &= \frac{2}{n} \sum_{t=1}^n z_t e^{-j f_i t} , \end{aligned}$$

we have

$$\begin{aligned} A(f_i) &= \left( \frac{2}{n} \right)^2 \sum_{t=1}^n z_t e^{-j f_i t} \sum_{t'=1}^n z_{t'} e^{+j f_i t'} \\ &= \left( \frac{2}{n} \right)^2 \sum_{t=1}^n \sum_{t'=1}^n z_t z_{t'} e^{-j f_i (t-t')} \\ &= \left( \frac{2}{n} \right)^2 \left| \sum_{t=1}^n z_t e^{-j f_i t} \right|^2 \\ &= \frac{2}{n} I_n^* (f_i) . \end{aligned}$$

Hence

$$I_n^* (f_i) = \frac{n}{2} A(f_i) = \frac{n}{2} (a_i^2 + b_i^2) .$$

is an alternative expression for the periodogram. The quantity  $I_n^* (f_i)$  is the sum of squares attributable to the coefficients  $a_i$  and  $b_i$  and so the  $I_n^* (f_i)$  can be used to construct an analysis of variance table to test their significance. Box and Jenkins illustrate this method. Alternatively, Wilks shows how to test for significance of the largest periodogram component.

## REFERENCES

B-1. Box, G.E.P., and Jenkins, G.M., *Time Series Analysis, Forecasting, and Control*, Holden-Day, Inc., San Francisco, Calif., 1970.

B-2. Wilks, S.S., *Mathematical Statistics*, John Wiley and Sons, New York, 1962.

## APPENDIX C. SPECTRAL ANALYSIS

This Appendix will present a description of the statistic used to estimate the spectrum of a time series, and a test for “whiteness” i.e., of deviation of the spectrum from a constant line.

### 1. Definition of the Spectrum

The power spectrum of the discrete time series  $\{a_t\}$  is defined as

$$p(f) = 2 \left( \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k \cos \left( \frac{\pi f k}{L} \right) \right) \quad 0 \leq f \leq L,$$

i.e., it is the Fourier cosine transform of the theoretical covariance function.

It follows that

$$\gamma_k = \frac{1}{2L} \int_0^L p(f) \cos \frac{k\pi f}{L} df, \quad k = 0, 1, 2, \dots,$$

i.e.,  $\gamma_k$  is the Fourier transform of the power spectrum. Choosing  $L = \pi$  we have

$$p(f) = 2 \left( \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k \cos fk \right), \quad 0 \leq f \leq \pi,$$

and

$$\gamma_k = \frac{1}{2\pi} \int_0^{\pi} p(f) \cos(kf) df, \quad k = 0, 1, 2, \dots$$

The value  $L = \frac{1}{2}$  is also a popular choice. It is a somewhat more natural choice, since the frequency is then measured in units of cycles per second, rather than radians per second. The spectrum is defined over a finite range (e.g.,  $-\frac{1}{2}, \frac{1}{2}$ ) for discrete (lattice) processes since any sine curve of frequency  $> \frac{1}{2}$  is indistinguishable from some sine curve with frequency  $< \frac{1}{2}$  on a lattice of unit time period spacing.

The spectral density function is defined by

$$\begin{aligned} g(f) &= \frac{p(f)}{2\pi\sigma^2} \\ &= \frac{1}{\pi} \left[ 1 + 2 \sum_{k=1}^{\infty} p_k \cos fk \right] \quad 0 \leq f \leq \pi. \end{aligned}$$

Since

$$\int_0^{\pi} g(f) df = 1,$$

and  $g(f)$  is positive, it has the properties of a probability density function.

## 2. Estimates of the Spectrum

### a. Raw Periodogram

A natural (but poor) estimate of the power spectrum is the periodogram:

$$I_n(f) = 2 \sum_{k=1-n}^{n-1} c_k e^{-ifk}$$

$$= 2 \left( c_0 + 2 \sum_{k=1}^{n-1} c_k \cos fk \right) \quad 0 \leq f \leq \pi ,$$

where

$$c_k = \frac{1}{n-k} \sum_{t=1}^{n-k} (a_t - \bar{a})(a_{t+k} - \bar{a}) .$$

An alternative definition of the periodogram is

$$I_n^*(f) = \frac{2}{n} \left[ \sum_{t=1}^n (x_t - \bar{x}) e^{-ift} \right]^2$$

$$= 2 \sum_{k=1-n}^{n-1} \left( 1 - \frac{|k|}{n} \right) c_k e^{-ifk}$$

$$= 2 \left[ c_0 + 2 \sum_{k=1}^{n-1} \left( 1 - \frac{k}{n} \right) c_k \cos fk \right]$$

i.e., there is an additional factor  $(1 - k/n)$  introduced. The above estimate is slightly better than the one given earlier, but the asymptotic properties are the same. The periodogram thus has the same form as the power spectrum, with the estimated covariances substituted for the theoretical covariances. That is, it is the Fourier cosine transform of the sample autocovariance function. (It is also seen to be the square of the Fourier transform of the data.) It is a poor estimate of the power spectrum, since its variance (for fixed  $f$ ) does not decrease to zero as the sample size ( $n$ ) increases, i.e., it is not a *consistent* estimate.

### b. Smoothed Periodogram

It can be shown that the modified estimate

$$p_n(f) = 2 \left( c_0 + 2 \sum_{k=1}^{n-1} \lambda_k c_k \cos fk \right) \quad 0 \leq f \leq \pi ,$$

is consistent (variance decreases to zero as  $n$  increases), using the appropriate values for the weights ( $\lambda$ 's). This estimate is called the “smoothed” periodogram, as opposed to the “raw” periodogram. Such weighted estimates are equivalent to estimates of the form

$$p_n(f) = \int_0^\pi w(f-h) I_n(h) dh$$

i.e., a weighted average of the periodogram over some band of frequencies. Thus we do not have a consistent estimate of the power spectrum at a single frequency; we do have, however, a consistent

estimate of a smoothed power spectrum. It is noted, however, that the integral of the raw periodogram is a consistent estimate of the spectral distribution function:

$$\int_0^f g(f) df.$$

The weight function  $\lambda_k$  is called a lag window; the weight function  $w(f)$  is called a spectral window. The pair  $(\lambda_k, w(f))$  of corresponding weights is called a window pair; they are Fourier transforms of each other.

The weighting can also be applied to the data itself,

$$p_n(f) = \frac{2}{n} \left| \sum_{t=1}^n \mu_t (x_t - \bar{x}) e^{-ift} \right|^2$$

in which case the weight function  $\mu_t$  is called a data window. The lag window, which is equivalent (in an expected value sense) to this data window, is

$$\lambda_k = \frac{1}{n-k} \sum_{t=1}^{n-k} \mu_t \mu_{t+k},$$

since

$$\text{cov}(\mu_t (x_t - \bar{x})) = \lambda_k c_k.$$

The Fourier transform,  $v(f)$ , of  $\mu_t$  is called the frequency window, and the equivalent spectral window (Fourier transform of  $\lambda_k$ ) is

$$w(f) = \frac{1}{n-k} |v(f)|^2.$$

Blackman and Tukey (Reference C-1) present a thorough discussion of smoothing. As Box and Jenkins note, the impracticability of estimating the spectral density function at an exact frequency is analogous to that of estimating an ordinary density function at an exact point using a histogram. By choosing suitably large intervals, the histogram successfully estimates the probability mass in the intervals; if the interval is too small, the estimate oscillates wildly along the axis, and from sample to sample. Similarly, by choosing a suitably broad spectral window, the smoothed periodogram successfully estimates the spectral mass over the window; if the window is too small, the variance of the corresponding estimate is so large that the estimate is useless.

### c. Weight Functions

Bartlett introduced the weights

$$\lambda_k = \begin{cases} 1 - k/m_n & \text{if } k \leq m_n \\ 0 & \text{if } k > m_n \end{cases}$$

where the parameter  $m_n$  is an integer less than  $n$  such that  $m_n \rightarrow \infty$  and  $m_n/n \rightarrow 0$  as  $n \rightarrow \infty$  (say  $\sim n^{1/3}$ ).

A number of other weighting functions have been considered for use with  $I_n^*(f)$ . Daniell introduced the weights

$$\lambda_k = \frac{\sin \pi k / 2m_n}{\pi k / 2m_n}, \quad m_n \leq n - 1.$$

These weights require computation of all  $n - 1$  covariances, unfortunately. (Once again,  $m_n \rightarrow \infty$  and  $m_n/n \rightarrow 0$  as  $n \rightarrow \infty$ ).

Approximations to the preceding weight function are given by Hanning:

$$\lambda_k = .54 + .46 \cos(\pi k / m_n)$$

and by von Hann:

$$\lambda_k = \frac{1}{2} (1 + \cos(\pi k / m_n)) .$$

Perhaps the simplest weights are

$$\lambda_k = \begin{cases} 1 & \text{if } k \leq m_n \\ 0 & \text{if } k > m_n \end{cases} ,$$

resulting in a “truncated” periodogram. This estimate has the unpleasant property, however, that it can be negative.

### 3. White Noise

A sequence  $\{a_i\}$  of uncorrelated random variables having constant mean and variance is called a “white noise” sequence. A white noise sequence is hence one that has zero autocorrelation function, and constant spectral density function, equal to  $1/\pi = .318$ . A test for zero autocorrelation function was given in Appendix A, assuming normality of the  $a_i$ 's.

### 4. A Test for Whiteness

A test for constant spectral density function is given by Grenander and Rosenblatt (Reference C-2). We have

$$\begin{aligned} & \lim_{n \rightarrow \infty} \Pr \left\{ \max_{0 \leq f \leq \pi} \sqrt{n} |F_n^*(f) - F^*(f)| \leq \alpha \right\} \\ &= \sum_{k=-\infty}^{\infty} (-1)^k [\phi((2k+1)\alpha/x) - \phi((2k-1)\alpha/x)] = \Delta(\alpha/x) \end{aligned}$$

where  $\phi(u)$  is the normal distribution function and

$$x = \sqrt{2\pi G(\pi)}$$

where  $G(\pi) = \frac{1}{4\pi} \sum_{-\infty}^{\infty} c_i^2$  .

The quantity  $F^*(f)$  is the un-normalized spectral distribution function,

$$\begin{aligned} F^*(f) &= \int_0^f p^{**}(f') df' \\ &= \frac{1}{2\pi} \left( \gamma_0 f + 2 \sum_{k=1}^{\infty} \gamma_k \frac{\sin kf}{k} \right), \quad 0 \leq f \leq \pi \end{aligned}$$

where the following definition is used for the spectrum:

$$p^{**}(f) = \frac{1}{2\pi} \left( \gamma_o + 2 \sum_{k=1}^{\infty} \gamma_k \cos fk \right), \quad 0 \leq f \leq \pi.$$

Note that

$$F^*(f) = \frac{\gamma_o f}{2\pi} = \frac{\sigma^2 f}{2\pi},$$

and that in particular,

$$F^*(\pi) = \frac{\gamma_o}{2} = \frac{\sigma^2}{2}.$$

The quantity  $G(\pi)$  is estimated by

$$G_n(\pi) = \frac{1}{4\pi} \left( c_0^2 + 2 \sum_{k=1}^{m_n} c_k^2 \left( 1 - \frac{k}{m_n} \right) \right)$$

where  $m_n = an^b$ ,  $0 \leq a \leq 1$ ,  $b > 0$ . The quantity  $F_n^*(f)$  is the estimate of  $F^*(\lambda)$  defined by

$$F_n^*(f) = \frac{1}{2\pi} \left( c_0 f + 2 \sum_{k=1}^{m_n} c_k \left( 1 - \frac{k}{m_n} \right) \frac{\sin(kf)}{k} \right),$$

i.e., the integral of the Bartlett smoothed periodogram.

For a 95% level of significance test of the hypothesis that the spectral density function is constant, we determine  $\alpha_{.95} = \Delta^{-1}(.95)/x$ , and reject the hypothesis if

$$D = \max_{0 \leq f \leq \pi} \sqrt{n} \left| F_n^*(f) - \frac{\sigma^2 f}{2\pi} \right| > \alpha_{.95}.$$

This test is not very satisfactory, however, since  $\sigma^2$  is usually unknown. Replacing  $\sigma^2$  by an estimate results in accepting the null hypothesis more often than we should.

Since estimates of only  $n-1$  autocorrelations are available, estimates of  $n-1$  different spectral mass densities contain all of the autocorrelation information of the sample, even though the estimate spectrum is plotted as a continuous curve. In fact, Hannan (Reference C-3) shows that the raw periodogram estimates

$$I_n^* \left( \frac{2\pi k}{n} \right) \quad (k = 1, 2, \dots, (n-1)/2)$$

are uncorrelated for normal processes.

Such is not the case, however, for the smoothed estimates  $p_n(k\pi/n)$ . The weight function  $w(f)$  used for smoothing the periodogram overlaps for adjacent estimated spectral masses (frequency distance  $\pi/m_n [= \pi/(n-1)]$  apart), resulting in high correlation between them. There is little overlap for next-adjacent estimates (distance  $2\pi/(n-1)$  apart), however, and they are hence essentially uncorrelated. The "resolution" of the smoothed periodogram (minimum frequency distances between essentially uncorrelated spectral mass estimates) is hence  $2\pi/(n-1)$ . See Blackman and Tukey (Reference C-1) or Hannan (Reference C-3) for further discussion of this point.

We shall describe a test for whiteness that does not require knowledge of the true variance ( $\sigma^2$ ) of the process. It can be shown that, under the hypothesis of whiteness, each of the quantities

$$I_n(f_k), \quad f_k = 2\pi k/(n-1), \quad k = 1, \dots, (n-1)/2$$

has the same distribution. A test of whether or not the spectral distribution function

$$F_n(f_k) = \frac{1}{(n-1)\hat{\sigma}^2} \sum_{i=1}^k I_n(f_i)$$

differs significantly from a white noise spectral distribution function

$$F_w(f_k) = 2\pi k / (n - 1)$$

can be tested using the Kolmogoroff-Smirnov test for goodness of fit. The test statistic is

$$\max_k |F_n(f_k) - F_w(f_k)|$$

and the 100p% critical value for the test statistic is (approximately)  $\lambda_p ((n - 1) / 2 - 1)^{-1/2}$  where  $\lambda_p$  (the  $p$ -th percentile point of the distribution of the test statistic) equals 1.36 for  $p = .95$  and 1.02 for  $p = .75$ . Reference C-4 may be consulted for additional details concerning the preceding test.

#### REFERENCES

C-1. Blackman, R.B., and Tukey, J.W., *The Measurement of Power Spectra*, Dover Publications, New York, 1959.

C-2. Grenander, U., and Rosenblatt, M., *Statistical Analysis of Stationary Time Series*, John Wiley and Sons, New York, 1957.

C-3. Hannan, E.J., *Time Series Analysis*, John Wiley and Sons, New York, 1960.

C-4. Jenkins, G.M. and Watts, D.G., *Spectral Analysis and its Applications*, Holden-Day, San Francisco, 1968.

## APPENDIX D. LINEAR DIFFERENCE EQUATIONS

As indicated in the text, the difference equation solution of the autoregressive part of the time series model is the median value of the process. This Appendix presents a fundamental result from the theory of linear difference equations, namely the solution of a linear second order difference equation.

Consider the difference equation

$$z_t - \phi_1 z_{t-1} - \phi_2 z_{t-2} = 0$$

where  $\phi_2 \neq 0$ . Let  $m_1$  and  $m_2$  denote the two roots of the corresponding indicial (auxiliary, characteristic) equation:

$$x^2 + \phi_1 x + \phi_2 = 0.$$

The general solution of the above difference equation is given by

$$z_t = c_1 m_1^t + c_2 m_2^t$$

if  $m_1$  and  $m_2$  are real and unequal, by

$$z_t = (c_1 + c_2 t) m_1^t$$

if  $m_1 = m_2$ , and by

$$z_t = c_1 r^t \cos(k\theta + c_2)$$

if  $m_1$  and  $m_2$  are the complex conjugates  $r(\cos \theta \pm i \sin \theta)$ .

The general solution to the difference equation

$$z_t - \phi_1 z_{t-1} - \phi_2 z_{t-2} = e_t$$

is given by

$z_t =$  general solution to the homogenous equation  
(obtained by setting  $e_t = 0$  in the above)  
+ particular solution of the complete equation.

If the roots  $m_1$  and  $m_2$  are unequal, then a particular solution of the complete equation is

$$u_t = \frac{m_1}{m_1 - m_2} u_{1t} - \frac{m_2}{m_1 - m_2} u_{2t}$$

where

$$u_{it} = e_t + m_i e_{t-1} + m_i^2 e_{t-2} + \dots + m_i^t e_0.$$

If  $m_1 = m_2 = m$ , then a particular solution of the complete equation is

$$v_t = e_t + (2m) e_{t-1} + (3m^2) e_{t-2} + \dots + ((t+1)m^t) e_0.$$

The constants of the general solution are chosen to satisfy the initial conditions

$$z_t = z_{(0)} \quad \text{for } t = 0$$

$$z_t = z_{(1)} \quad \text{for } t = 1.$$

Goldberg (Reference D-1) provides a readable introduction to difference equations.

### REFERENCES

D-1. Goldberg, Samuel, *Introduction to Difference Equations*, Science Editions, Inc., John Wiley and Sons, Inc., New York, 1961.

# APPENDIX E. LEAST-SQUARES AND MAXIMUM LIKELIHOOD ESTIMATION

## 1. Introduction

This Appendix shows the derivation of the least-squares estimates of the parameters of the model

$$\Phi(B)z_t = \Theta(B)a_t$$

and gives the formula for the likelihood function in the case of normally distributed  $a_t$ 's.

The analysis in the case of multiplicative seasonal components is similar and will not be described here. It is noted that if we have a nonseasonal pure autoregressive model, the model is what is known as a *linear model*, and the estimation is straightforward. If any  $\theta$ 's are present, however, or if there is more than one  $\phi$  polynomial, then the model is not linear, and iterative procedures are employed to find the least-squares estimates.

## 2. Least-Squares Estimates

### a. Pure Autoregressive Model

In the case of the pure autoregressive model, the least-squares estimates are easy to compute. We have

$$\Phi(B)z_t = a_t$$

or

$$z_t = \underline{z}_t' \underline{\phi} + a_t$$

where

$$\underline{z}_t = (z_{t-1}, z_{t-2}, \dots, z_{t-p})$$

and

$$\underline{\phi}' = (\phi_1, \phi_2, \dots, \phi_p) .$$

Suppose we have  $n$  observations. Denoting the vector of the  $n-p$  most recent observations by

$$\underline{z}' = (z_n, z_{n-1}, \dots, z_{p+1})$$

and the error vector (i.e., vector of random components) by

$$\underline{a}' = (a_n, a_{n-1}, \dots, a_{p+1})$$

we have

$$\underline{z} = Z \underline{\phi} + \underline{a}$$

where

$$Z = \begin{pmatrix} z_n' \\ z_{n-1}' \\ \vdots \\ z_{p+1}' \end{pmatrix} .$$

This model is linear in the  $\phi$ 's, and so it is easy to calculate the least-squares estimates of the  $\phi$ 's, given by the values which minimize the residual sum of squares:

$$\begin{aligned} S &= (\underline{z} - Z\underline{\phi})' (\underline{z} - Z\underline{\phi}) \\ &= (\underline{z}' - \underline{\phi}' Z') (\underline{z} - Z\underline{\phi}) \\ &= \underline{z}' \underline{z} - \underline{\phi}' Z' \underline{z} - \underline{z}' Z \underline{\phi} + \underline{\phi}' Z' Z \underline{\phi} \\ &= \underline{z}' \underline{z} - 2\underline{z}' Z \underline{\phi} + \underline{\phi}' Z' Z \underline{\phi} . \end{aligned}$$

Applying vector differentiation, we have

$$\begin{aligned} \frac{dS}{d\underline{\phi}} &= -2\underline{z}' Z + \underline{\phi}' Z' \frac{dZ\underline{\phi}}{d\underline{\phi}} + \underline{\phi}' Z' \frac{dZ\underline{\phi}}{d\underline{\phi}} \\ &= -2\underline{z}' Z + 2\underline{\phi}' Z' Z . \end{aligned}$$

Hence, setting  $\frac{dS}{d\underline{\phi}}$  equal to the null vector, we have

$$-2\underline{z}' Z + 2\underline{\hat{\phi}}' Z' = \underline{0}$$

or

$$\underline{\hat{\phi}}' Z' Z = \underline{z}' Z$$

or

$$Z' Z \underline{\hat{\phi}}' = Z' \underline{z}$$

or

$$\underline{\hat{\phi}} = (Z' Z)^{-1} Z' \underline{z}$$

as the least squares estimate of  $\underline{\phi}$ . The estimate of the variance-covariance matrix of  $\underline{\hat{\phi}}$  is

$$\Sigma = (Z' Z)^{-1} \hat{\sigma}^2$$

where  $\hat{\sigma}^2$ , the estimate of the variance of  $a_t$ , is given by

$$\begin{aligned} \hat{\sigma}^2 &= (\underline{z} - Z\underline{\hat{\phi}})' (\underline{z} - Z\underline{\hat{\phi}}) \\ &= \underline{z}' \underline{z} - \underline{z}' Z \underline{\hat{\phi}} . \end{aligned}$$

## b. Mixed Autoregressive Moving Average Model

We shall now determine the least-squares estimates in the case of the mixed autoregressive moving average model. Suppose that the zeros of  $\Theta(B)$  are outside the unit circle. We may then invert the process and write

$$a_t = \Theta^{-1}(B) \Phi(B) z_t$$

Considering  $a_t = a_t(\underline{\beta}, z_t, z_{t-1}, \dots)$  as a function of  $\underline{\beta}' = (\underline{\theta}', \underline{\phi}')$  and the  $z$ 's up to and including  $z_t$ , we expand  $a_t$  in a Taylor series about some trial value  $\underline{\beta}^0$ :

$$a_t \doteq a_t^0 - \sum_{i=1}^{p+q} (\beta_i - \beta_i^0) X_{it}^0$$

where

$$X_{it}^0 = - \left. \frac{\partial a_t}{\partial \beta_i} \right|_{\underline{\beta} = \underline{\beta}^0} .$$

Denoting:

$$\underline{d}^0 = \underline{\beta}^0 - \underline{\beta}^0 ,$$

$$\underline{a}' = (a_n, a_{n-1}, \dots, a_1) ,$$

and

$$\underline{a}^0 = (a_n^0, a_{n-1}^0, \dots, a_1^0) ,$$

and denoting the matrix whose  $(i, t)$ -th element is  $X_{it}^0$  by  $X$ , we may write the above as

$$\underline{a} = \underline{a}^0 - X \underline{d}^0$$

or

$$\underline{a}^0 = X \underline{d}^0 + \underline{a}$$

which is a linear model in the deviations  $\underline{d}^0$  of the trial value  $\underline{\beta}^0$  from the true value  $\underline{\beta}$ . As before, the least-squares estimates for  $\underline{d}^0$  are

$$\hat{\underline{d}}^0 = (X'X)^{-1} X' \underline{a}^0 .$$

We calculate the new trial value  $\underline{\beta}^1 = \underline{\beta}^0 + \hat{\underline{d}}^0$ , and repeat the above process to obtain  $\hat{\underline{d}}^1$  and hence  $\underline{\beta}^2 = \underline{\beta}^1 + \hat{\underline{d}}^1$ . This process is continued until  $\underline{\beta}^i$  converges. If a sufficiently bad initial trial value for  $\underline{\beta}^0$  is used, the iteration might not converge. Box and Jenkins suggest means for picking “reasonable” trial values. In practice, zero initial values are satisfactory.

For each iteration the values of the quantities  $a_t^0$  and  $X_{it}^0$  are calculated iteratively as follows.

Since

$$a_t^0 = \Theta^{0-1}(B)\Phi^0(B)z_t ,$$

we have

$$a_t^0 = \theta_1^0 a_{t-1} + \theta_2^0 a_{t-2} + \dots + \theta_q^0 a_{t-q} + z_t - \phi_1^0 z_{t-1} - \dots - \phi_p^0 z_{t-p}$$

since

$$v_{it}^0 = - \left. \frac{\partial a_t}{\partial \theta_i} \right|_{\underline{\beta} = \underline{\beta}^0} = - \Theta^{0-2}(B)B^i z_t = - \Theta^{0-1}(B)a_{t-i}^0 ,$$

we have

$$v_{it}^0 = \theta_1^0 u_{it-1}^0 + \theta_2^0 v_{it-2}^0 + \dots + \theta_q^0 v_{it-q}^0 - a_{t-i}^0 ;$$

since

$$u_{it}^0 = - \left. \frac{\partial a_t}{\partial \phi_i} \right|_{\underline{\beta} = \underline{\beta}^0} = \Theta^{0-1}(B)B^i z_t = \Phi^{0-1}(B)a_{t-i}^0 ,$$

we have

$$u_{it}^0 = \phi_1^0 u_{it-1}^0 + \phi_2^0 u_{it-2}^0 + \dots + \phi_q^0 u_{it-q}^0 + a_{t-i}^0 .$$

### 3. Maximum Likelihood Estimates

The above procedure for obtaining estimates of the  $\phi$ 's and  $\theta$ 's produces so-called "least-squares" estimates; i.e., the estimates are chosen so as to minimize the sum of squares of the deviations of the observations from their expectations.

An alternative procedure for obtaining estimates is to choose those values for the  $\phi$ 's and  $\theta$ 's that maximize the likelihood function (i.e., the probability function of  $a_1, \dots, a_n$ , considered as a function of the  $\phi$ 's and  $\theta$ 's). For the case in which the  $a$ 's are normally distributed, the maximum likelihood estimates are the same as the least-squares estimates. In addition, the maximum likelihood procedure is useful in that knowledge of the entire likelihood function (and not just its maximum) can provide insight into the behavior of the model as the parameters are varied from their maximizing values. This would enable one to determine more useful forms of the time series model which, statistically speaking, are about as good. Also, parameter redundancies caused by common factors in the  $\Phi$  and  $\Theta$  polynomials would be recognized (the maximum might tend to lie along a line).

Another reason for using the maximum likelihood method is that we are restricting our  $\phi$  parameters so that the zeros of  $\Phi(B)$  are outside the unit circle. The least-squares procedure is not designed to handle such a restriction. Using the maximum likelihood method, however, we may simply choose the parameter estimates corresponding to the maximum value of the likelihood function in the acceptable region.

Note that it is, of course, necessary to assume a distribution for the  $a_t$ 's if we wish to use the maximum likelihood method.

### 4. The Likelihood Function

As noted above, we assume normality of the random variation ( $a_t$ 's) of the time series. The probability density function of the  $a_t$ 's is hence

$$f(a_1', \dots, a_n') = \left( \frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp \left( \frac{-\sum_{t=1}^n (a_t' - \mu)^2}{2\sigma^2} \right)$$

where  $\mu = Ea_t'$  and  $\sigma^2 = \text{var } a_t'$ ,  $t = 1, \dots, n$ . For given values of the parameters  $\mu$  and  $\sigma^2$  the probability density function indicates the "likelihood" of an outcome  $a_1', \dots, a_n'$ . If the parameters are known, then a set of intuitively appealing estimates are those values for which  $f$  is maximized. The function  $f$ , considered as a function of the parameters, is called the *likelihood function*, and will be denoted by  $L$ . The maximizing parameter values are called *maximum likelihood estimates*. Maximum likelihood estimates can be shown to have desirable statistical properties. In the case of a normal distribution (as above), the least-squares estimates and the maximum likelihood estimates are identical.

The likelihood function is

$$L(\mu, \sigma^2) = \left( \frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp \left( \frac{-\sum_{t=1}^n (a_t' - \mu)^2}{2\sigma^2} \right)$$

It is convenient to work with the log-likelihood function,  $\ell = \ln L$  (a monotonic function of  $L$ ):

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^n (a_t' - \mu)^2 .$$

Let us denote  $a_t = a_t' - \mu$ ;  $a_t$  has zero expectation.

Note that in the time series situation, we cannot observe the  $a_t$ 's directly. Instead, we can only compute them, given values for the  $\phi$ 's and  $\theta$ 's

Since the  $a_t$ 's depend on the parameters ( $\phi$ 's and  $\theta$ 's) of the time series model and on previous observations, we write

$$a_t = a_t(\underline{\phi}, \underline{\theta} | \underline{S})$$

where  $\underline{\phi} = (\phi_1, \dots, \phi_p)$ ,  $\underline{\theta} = (\theta_1, \dots, \theta_q)$ , and  $\underline{z} = (z_1, \dots, z_n)$ .

Since  $a_t$  depends on  $\phi$  and  $\theta$ , so does  $\ell(\mu, \sigma^2)$ , and so we write  $\ell(\mu, \sigma^2, \phi, \theta)$  for completeness.

In general, our model is

$$\Phi(B)z_t = \Phi_p^*(B)(1-B)^d z_t = \Theta_q(B)a_t' = \Theta_q(B)(a_t + \mu) = \theta_0 + \Theta_q(B)a_t$$

where

$$\theta_0 = \Theta_q(B)\mu,$$

so that

$$a_t = \mu + \Theta^{-1}(B)\Phi^*(B)(1-B)^d z_t .$$

If all of the  $\theta$ 's are zero, the model is linear (in the coefficients of the  $z_t$ 's) and  $a_t$  involves only  $z_1, \dots, z_n$ , given values for the  $\phi$ 's and  $d$ :

$$a_t = \mu + \Phi^*(B)(1-B)^d z_t .$$

Calculation of the maximum likelihood estimates of the  $\phi$ 's,  $\mu$ , and  $\sigma^2$  is straightforward in this case:

$$\hat{\mu} = \frac{1}{n} \sum_{t=1}^n a_t'$$

$$\sigma^2 = \frac{1}{n} \sum_{t=1}^n (a_t' - \hat{\mu})^2$$

and the  $\phi$ 's as given previously. If any of the  $\theta$ 's are nonzero, however, the model is nonlinear and  $a_t$  depends on an infinite number of unobservable  $z_t$ 's (namely  $z_0, z_{-1}, z_{-2}, \dots$ ).

Writing the  $a_t$ 's recursively, we have

$$a_t = \theta_1 a_{t-1} + \theta_2 a_{t-2} + \dots + \theta_q a_{t-q} + \theta_0 + z_t - \phi_1 z_{t-1} - \phi_2 z_{t-2} - \dots - \phi_p z_{t-p} .$$

Hence, in addition to  $\mu, \sigma^2$ , the  $\theta$ 's and the  $\phi$ 's we must estimate  $a_0, a_{-1}, \dots, a_{1-q}$  and  $z_0, z_{-1}, \dots, z_{1-p}$ . With a large number of observations, however, it is justifiable simply to ignore  $a_1, a_2, \dots, a_p$ , and set  $a_t, p < t \leq q$  equal to zero. Then we need estimate only  $\mu, \sigma^2$ , the  $\theta$ 's and the  $\phi$ 's.

## APPENDIX F. TRANSFORMATION OF DATA

One of the underlying assumptions in the least-squares estimation procedure is that the variance of the  $a_t$ 's are homoscedastic, or constant for varying  $t$ . If this is not true, then the estimates are still unbiased, but their sampling variances are inflated. A Bartlett's  $M$ -test can be applied to test whether or not the variances of different series to be pooled for estimation purposes are heteroscedastic. In addition to variance differences from series to series, it may be that the variance of the same series changes over the length of the series. To enable efficient estimation, it is desirable to transform the data so that the variance is comparable for all the observations.

If  $m = E(X)$ , and  $\text{var } X = f(m)$ , then a transformation which will usually result in a random variable with constant variance is

$$Y = g(X) = \int^X \left[ \frac{1}{f(u)} \right]^{1/2} du \quad ,$$

since, by Taylor's theorem,

$$\begin{aligned} \text{var}(Y) &\doteq [g'(m)]^2 \text{var}(X) \\ &= \left[ \frac{d}{dX} \int^X \left[ \frac{1}{f(u)} \right]^{1/2} du \right]^2 \Big|_{X=m} f(m) \\ &= 1 \end{aligned}$$

provided suitable regularity conditions are satisfied. For example, if the magnitude of observation-to-observation variations in  $z_t$  appear to be proportional to the local mean level, the logarithmic transformation is appropriate, since, letting  $\text{SD}(X) = \mu$ , or  $\text{var}(X) = k^2 \mu^2$  we have

$$\begin{aligned} g(X) &= \int^X \left[ \frac{1}{k^2 \mu^2} \right]^{1/2} du \\ &= \int^X \frac{1}{k\mu} du \\ &= \frac{1}{k} \log X \quad . \end{aligned}$$

Note that transformation of the data has the effect that the estimates are "least-squares" with respect to the transformed scale, not the original scale.

### Forecasting in the Case of a Logarithmic Transformation

It is of interest to observe the effect of a data transformation on forecasts. Suppose that fluctuations in the original time series are proportional to the level of the time series, so that it is appropriate to apply a logarithmic transformation to the data. Let us denote the untransformed variate by  $z_t'$ . To forecast  $z_t'$ , we compute the least-squares forecast of the logarithm,  $z_t$ , of  $z_t'$  and exponentiate the result. Although, the forecast logarithm is the expected value (mean) of  $z_t$ , the exponentiated result is not the expected value of the original variate,  $z_t'$ . It is, however, the median of the distribution of  $z_t'$ , if the  $z_t'$ 's are normally distributed. The median value,  $e^{Ez_t}$ , of  $z_t'$  at time  $t$  is used rather than the mean value,  $Ee^{z_t}$ , to forecast  $z_t'$  at time  $t$  for the following reason. Suppose that  $z_t'$  obeys a lognormal distribution.

The median is considered more appropriate than the mean as a forecaster for a lognormal distribution, since the mean value of a lognormal distribution is unduly influenced by extremely improbable, extremely large values. For example, the mean can be thousands of times larger than the .99 percentile point of the distribution.

Note, however, that the median forecaster is biased low. This follows from Jensen's inequality:

$$e^{E X} \leq E e^X$$

where E is the expectation operator. It follows that

$$\exp(\text{least squares forecast log sales}) \leq \text{least squares forecast sales};$$

i.e., the forecaster is biased low. The bias increases as the variability ( $\text{var } a_t$ ) increases. In particular, if  $z_t'$  is lognormally distributed,

$$z_t' \sim LN(\mu, \sigma^2)$$

then

$$z_t = \ln z_t' \sim N(\mu, \sigma^2)$$

and

$$E e^{z_t} = E z_t' = e^{\mu + \sigma^2/2} > e^\mu = e^{E z_t} .$$

The bias is the factor  $\exp(\sigma^2/2)$  which increases as  $\sigma^2$  increases.

## APPENDIX G. EXPONENTIAL SMOOTHING

A procedure that has been widely used in forecasting is *exponential smoothing*. This Appendix will show that the use of exponential smoothing in fact corresponds to use of a particular model of the form

$$\Phi(B)z_t = \Theta(B)a_t \quad .$$

Suppose that

$$\Phi(B) = 1 - B = \nabla$$

and that

$$\Theta(B) = 1 - \alpha B \quad ,$$

so that our model is

$$\nabla z_t = z_t - z_{t-1} = a_t - \alpha a_{t-1} \quad . \quad (G-1)$$

The lead-one least-squares forecaster associated with this model is

$$\hat{z}_t(1) = z_t - \alpha \hat{a}_t$$

where  $\hat{a}_t = z_t - \hat{z}_{t-1}(1)$ , and  $\hat{z}_t(1)$  denotes the one-step-ahead forecast made from time  $t$ . We may hence write

$$\hat{z}_t(1) = z_t - \alpha (z_t - \hat{z}_{t-1}(1)) = (1 - \alpha)z_t + \alpha \hat{z}_{t-1}(1) \quad . \quad (G-2)$$

From this expression, we can obtain an alternative expression for the least-squares forecaster:

$$\begin{aligned} \hat{z}_t(1) &= (1 - \alpha)z_t + \alpha[(1 - \alpha)z_{t-1} + \alpha \hat{z}_{t-2}(1)] \\ &\dots = (1 - \alpha)[z_t + \alpha z_{t-1} + \alpha^2 z_{t-2} + \dots] \quad , \quad (G-3) \end{aligned}$$

which would readily have been derived from representing the process as

$$(1 - \alpha B)^{-1} \nabla z_t = a_t$$

or

$$\nabla z_t + \alpha \nabla z_{t-1} + \alpha^2 \nabla^2 z_{t-2} + \dots = a_t$$

or

$$z_t - z_{t-1} + \alpha z_{t-1} - \alpha z_{t-2} + \alpha^2 z_{t-2} - \alpha^2 z_{t-3} + \dots = a_t$$

or

$$z_t = (1 - \alpha)(z_{t-1} + \alpha z_{t-2} + \alpha^2 z_{t-3} + \dots) + a_t \quad ,$$

so that

$$z_{t+1} = (1 - \alpha)(z_t + \alpha z_{t-1} + \alpha^2 z_{t-2} + \dots) + a_{t+1} \quad .$$

The expression (G-3) is called an exponentially weighted moving average, and use of this forecaster is called (single) exponential smoothing. Although single exponential smoothing has been uncritically applied to forecast sales (having only changes in level), it is obviously appropriate only when the model (G-1) is a valid statistical representation of the sales time series. The program *TIMES* provides the means for determining whether or not the model (G-1) is appropriate, and for determining the correct model, if it is not.