SOME PROBLEMS IN THE SPECTRAL ANALYSIS OF TIME SERIES WITH MISSING VALUES

by

PETER BLOOMFIELD

Thesis submitted for the degree of Ph.D.
in the University of London

Imperial College, 1970

ABSTRACT

The problem considered is that of estimating the spectral density function of a time series from an incomplete set of observations. A general type of estimate, termed the natural estimate, is introduced.

The missing data is represented as follows. The series is completed by inserting zeros where values are missing. A new series, the amplitude modulating process, is constructed, with the value one where an observation on the original series is made, and zero where an observation is missed. The completed set of observations can then be regarded as the full series multiplied by the amplitude modulating process. For this reason, it is called the amplitude modulated process.

Several probabilistic structures for an amplitude modulating process are examined, and the asymptotic variances of two natural estimates are found in each case. The first of these structures is one in which the amplitude modulating process is weakly stationary, and has an absolutely continuous spectrum. In this case, it is found that the asymptotic variances of the two natural estimates considered depend only on the second-order properties of the amplitude modulating process. The remaining structures involve periodic effects, and the corresponding asymptotic variances are found to be more complicated.

A computer simulation study was carried out to assess the

closeness of the approximations given by these asymptotic results. The conclusion is drawn that the degree of approximation is comparable with that of the approximations used when the data is complete.

ACKNOWLEDGEMENTS

I am grateful to my supervisor, Professor David R. Cox, for his encouragement and advice, and for the original suggestion which led to this work. I would also like to thank Dr. hilton D. Miller, who initially supervised my work.

My thanks are due to Dr. Agnes M. Herzberg for helpful comments on the final version of this thesis.

During most of my research, I was supported by the Science Research Council, to whom I am also grateful.

CONTENTS

Title Page ..... Page 1

Abstract 2

Acknowledgements 4

Contents 5

Chapter 1. General Introduction 7

1.1 The problem 7

1.2 Preliminary details 9

1.3 Natural estimates and the amplitude modulating ~~process~~ function 15

Chapter 2. A Stationary Random Amplitude Modulating Process 18

2.1 Assumptions 18

2.2 The amplitude modulated process ... 22

2.3 An estimate of $g_X(x)$ 25

2.4 A more useful estimate 30

2.5 Two specific cases 40

2.6 Nonstationary deletion of data 45

Chapter 3. Periodically Missing Observations .. 57

3.1 Introduction 57

3.2 Periodically missing observations . 59

3.3 The mixed model 68

CONTENTS (Contd.)

Chapter 4. A More General Periodic Model ..... Page 73

4.1 Assumptions 73

4.2 The general result 75

4.3 Independent omissions 81

4.4 Delayed stationary omissions 83

4.5 Applications 85

4.6 A stationary model 87

Chapter 5. Other Estimates of the Spectral Density Function 89

5.1 Introduction 89

5.2 Neave's "alias-improved" estimate 89

5.3 Replacement of missing data 92

Chapter 6. Multivariate Data 99

6.1 Introduction 99

6.2 A multivariate stationary amplitude modulating process 102

6.3 Some special cases 105

Chapter 7. A Simulation Study 107

References 117

CHAPTER 1. GENERAL INTRODUCTION

1.1 The Problem

Spectral analysis is a branch of time series analysis which is concerned with the identification of dominant frequencies in a time series. More explicitly it involves the estimation of spectral distribution and density functions. In the past, the use of spectral analysis has been restricted by theoretical and computational difficulties.

The theoretical difficulties have now been largely overcome. A brief summary of the mathematical results involved is given in Section 1.2 ; for a fuller discussion, see Rosenblatt (1959) or Parzen (1961).

The computational difficulty is the large amount of computation involved in the spectral analysis of a time series of only moderate length. This has become less important with the development of the modern high-speed digital computer. The rediscovery of the Fast Fourier Transform has also helped, by reducing the actual amount of computation needed (Cooley, Lewis and Welch, 1967).

The problems caused by incomplete data have received little attention. These problems are possibly more serious in time series analysis than in other branches of statistics, as there will often be only one realisation of a series available for analysis.

The first work on such problems was by Jones (1962),

who considered data in which a block of observations is periodically missed. For example, the data could have been recorded only on weekdays, with all weekends missed. Jones produced a consistent estimate of the spectral density function from such data, and derived its asymptotic variance.

Parzen (1962) developed the theory of asymptotically stationary processes, and applied it (Parzen, 1963) to to missing data problems. This generalised Jones's result to the situation in which observations are missed in a general periodic pattern.

A rather different way in which data can be incomplete is if some random mechanism applies, such as the random failure of a recording apparatus. The only previous attempt to handle such data is that of Scheinok (1965). He, however, assumed the rather simple model in which observations are made or not according to the outcomes of a sequence of Bernoulli trials. This is rather restrictive. For example, it would exclude the random failure of some apparatus resulting in the omission of a string of observations.

Some more general models for the random deletion of data are discussed. The first is a model in which a certain associated sequence is a stationary process with an absolutely continuous spectrum. Then models with a periodic structure are considered. Essentially the same estimate of the spectral density function is used throughout, and

its asymptotic variance is derived in each case.

In a recent paper, Neave (1970) discusses the spectral analysis of 'initially scarce' data. Here the sampling interval has been changed by an integral factor at some point in the record. For example, one might have daily data which was originally recorded on only one day each week. As this can be regarded as a missing data problem, a brief discussion is included.

1.2 Preliminary Details

The spectral analysis of a time series is based on the assumption that it is a realisation of a stochastic process. Suppose that $\{X(t): t = 0, \pm 1, \ldots\}$ is a real stochastic process with stationary first, second and fourth moments. Subsequently this will be indicated by " $X(t)$ is weakly stationary". This means that $E\{X(t)\}$ is a constant, which is assumed to be zero. It also means that

$$E\{X(t)X(t+r)\} = R_X(r)$$

is independent of t. The sequence $\{R_X(r): r = 0, \pm 1, \ldots\}$ is the auto-covariance function of $\{X(t)\}$. It is clear that $R_X(-r) = R_X(r)$.

It is a standard result that $R_X(r)$ may be written

$$R_X(r) = \int_{-\pi}^{\pi} e^{-irx} \, dG_X(x) \, , \qquad (1.1)$$

where $G_X(x)$ is the spectral distribution function and is nondecreasing. Note that the variance of $X(t)$ is

$$R_X(0) = \int_{-\pi}^{\pi} dG_X(x) = G_X(\pi) - G_X(-\pi).$$

Therefore, $G_X(x)$ is not, in general, a probability distribution function. Sometimes $G_X(x)$ is normalised, but this will not be done here.

When $G_X(x)$ is absolutely continuous, (1.1) may be written

$$R_X(r) = \int_{-\pi}^{\pi} e^{-irx} g_X(x)\,dx, \tag{1.2}$$

where $g_X(x) = G_X'(x)$ is the spectral density function (the prime denotes differentiation). The function $g_X(x)$ is normally only defined on the interval $(-\pi,\pi)$. It will be convenient to define it as a function of period 2π on the whole real line.

Under suitable conditions, (1.2) may be inverted to give

$$g_X(x) = (2\pi)^{-1} \sum_{r=-\infty}^{\infty} R_X(r) e^{irx}. \tag{1.3}$$

One such condition is

$$\sum |rR_X(r)| < \infty, \tag{1.4}$$

which ensures that $g_X(x)$ is continuous and continuously differentiable.

The stationarity of the fourth moments of $\{X(t)\}$ means that

$$E\{X(t)X(t+q)X(t+r)X(t+s)\} = P_X(q,r,s)$$

is independent of t. The fourth cumulant function of $\{X(t)\}$ is

$$Q_X(q,r,s) = P_X(q,r,s) - R_X(q)R_X(r-s) - R_X(r)R_X(s-q) - R_X(s)R_X(q-r),$$

which vanishes for a Gaussian process. Thus $Q_X(q,r,s)$ measures the difference between the fourth moments of $\{X(t)\}$ and those of a Gaussian process with the same auto-covariance function. These differences are assumed to be so small that

$$\sum_{q,r,s} |Q_X(q,r,s)| = Q_X < \infty. \qquad (1.5)$$

Given a stretch of data $\{X(t): t = 1, \ldots, n\}$, its periodogram is defined by

$$I_X(x) = (2\pi n)^{-1} \left| \sum_{t=1}^{n} X(t)e^{itx} \right|^2 \qquad (1.6)$$

$$= (2\pi)^{-1} \sum_{|r|<n} c_X(r)e^{irx},$$

where

$$c_X(r) = n^{-1} \sum_{t=1}^{n-|r|} X(t)X(t+|r|) \quad . \tag{1.7}$$

Note that $I_X(x)$ and $c_X(r)$ depend on n as well as on X and x or r. This dependence is omitted in the notation.

From (1.4), it may be shown that

$$E\{I_X(x)\} = g_X(x) + O(n^{-1}) \tag{1.8}$$

uniformly in x, as $n \to \infty$. If (1.5) holds, it may be shown that

$$\mathrm{Cov}\{I_X(x_1), I_X(x_2)\} = g_X(x_1)^2 \left[\left\{ \frac{\sin \frac{1}{2}n(x_1-x_2)}{n \sin \frac{1}{2}(x_1-x_2)} \right\}^2 + \left\{ \frac{\sin \frac{1}{2}n(x_1+x_2)}{n \sin \frac{1}{2}(x_1+x_2)} \right\}^2 \right] + O(n^{-1})$$

uniformly in x_1 and x_2, as $n \to \infty$ (Brillinger & Rosenblatt, 1967).

It follows that

$$\mathrm{Var}\{I_X(x)\} \to g_X(x)^2$$

as $n \to \infty$, so that although $I_X(x)$ is an asymptotically unbiased estimate of $g_X(x)$, it is not consistent. However, if $x_1 \neq \pm x_2$, it also follows that

$$\mathrm{Cov}\{I_X(x_1), I_X(x_2)\} = O(n^{-1})$$

as $n \to \infty$, whence even adjacent values of $I_X(x)$ tend to be uncorrelated. This suggests that an estimate with a smaller asymptotic variance could be constructed by averag-

ing over adjacent values of x. This gives rise to a smoothed estimate

$$g_X^{(1)}(x) = \int_{-\pi}^{\pi} I_X(y) w_n(y;x)\,dy \ , \tag{1.9}$$

where $w_n(y;x)$, for fixed x, is a weighting function with its weight concentrated around x.

The following conditions are imposed on $w_n(y;x)$:

$$\int_{-\pi}^{\pi} w_n(y;x)\,dy = 1 \ ; \tag{1.10}$$

$$\left.\begin{aligned} &2\pi\int_{-\pi}^{\pi} w_n(y;x)^2\,dy = W_n < \infty \ , \\ &2\pi\int_{-\pi}^{\pi} w_n(y+x_1;x_2)\,w_n(y;x_2-x_1)\,dy \sim W_n \\ \text{and}\quad &2\pi\int_{-\pi}^{\pi} w_n(y+x_1;x_2)\,w_n(-y;x_1-x_2)\,dy \sim W_n \ ; \end{aligned}\right\} \tag{1.11}$$

for any $\varepsilon > 0$, $w_n(y;x) \to 0$ as $n \to \infty$, uniformly in $|y - x| > \varepsilon$; (1.12)

if

$$w_r^{(n)}(x) = \int_{-\pi}^{\pi} e^{iry}\, w_n(y;x)\,dy \ , \tag{1.13}$$

then $|w_r^{(n)}(x)| \leq M < \infty$ for all r, n and x ;

$$\frac{\sum_{|r|<n} (1 - n^{-1}|r|)\,|w_r^{(n)}(x)|^2}{\sum_{r=-\infty}^{\infty} |w_r^{(n)}(x)|^2} \to 1 \tag{1.14}$$

as $n \to \infty$.

Note that (1.12) implies that $W_n \to \infty$ as $n \to \infty$. Henceforth, the dependence of $w_n(y;x)$ and $w_r^{(n)}(x)$

on n will be omitted in the notation.

Under conditions (1.4), (1.5) and (1.10) to (1.14), it may be shown that

$$\mathrm{Var}\left\{g_X^{(1)}(x)\right\} = n^{-1} W_n g_X(x)^2 \, d(x) + o(n^{-1} W_n) \quad , \qquad (1.15)$$

where $d_0(x) = 1$, $x = 0$, and $d_0(x) = 0$ otherwise, and $d(x) = 1 + d_0(x-\pi) + d_0(x) + d_0(x+\pi)$, and that

$$\mathrm{Cov}\left\{g_X^{(1)}(x_1) \ , \ g_X^{(1)}(x_2)\right\} = o(n^{-1} W_n) \qquad (1.16)$$

if $x_1 \neq \pm x_2$ (Rosenblatt, 1959). These results may also be inferred from a result in Section 3.2.

Now (1.12) is sufficient for $g_X^{(1)}(x)$ to be asymptotically unbiased. Thus, if the weighting functions are chosen so that $n^{-1} W_n \to 0$ as $n \to \infty$, it follows that $g_X^{(1)}(x)$ is a consistent estimate of $g_X(x)$. It is easy to construct sequences of functions with the required properties. This is discussed by Rosenblatt (1959).

Note that, in general, $g_X^{(1)}(x)$ is a biased estimate of $g_X(x)$, since it contains estimates of at most n autocovariances. Note also that since $W_n \to \infty$ as $n \to \infty$, $\mathrm{Var}\left\{g_X^{(1)}(x)\right\}$ tends to zero more slowly than n^{-1} .

Another way to write (1.9), using (1.13), is

$$g_X^{(1)}(x) = (2\pi)^{-1} \sum_{|r|<n} c_X(r) w_r(x) \ . \qquad (1.17)$$

In order to avoid the complications arising in the integration of random functions, the integral (1.9) will be defined

by the sum (1.17). As in (1.6), this may also be written

$$g_X^{(1)}(x) = (2\pi n)^{-1} \sum_{s,t=1}^{n} X(s)X(t)w_{s-t}(x) \ .$$

The most convenient of these forms will be used in each situation.

The function $g_X(x)$ is nonnegative, and it is sometimes required that its estimate $g_X^{(1)}(x)$ should be so as well. Now $I_X(x)$ is clearly nonnegative, and so if $w(y;x) \geq 0$, then $g_X^{(1)}(x)$ will also be positive. Unfortunately, none of the estimates to be discussed in missing data situations share this property, apart from one suggested by Neave (1970).

In this section it has been assumed that the mean of the observed process is zero. However, this is no restriction, in the following sense. If the mean is nonzero, it may be estimated and subtracted from the series. The effects of this on the mean and variance of $g_X^{(1)}(x)$ are both $O(n^{-1})$, so that the asymptotic mean and variance are unaltered. The same is true if the mean value has a more complex linear structure (Parzen, 1957).

1.3 Natural Estimates and the Amplitude Modulating Function

A comparison of the expression (1.17) for $g_X^{(1)}(x)$ with the expression (1.3) for $g_X(x)$ shows that the estimate can be described as follows. The auto-covariance $R_X(r)$ has

been replaced by $c_X(r)$, which under present assumptions is a consistent estimate of $R_X(r)$, and $\exp(irx)$ has been replaced by $w_r(x)$, which for fixed r approaches $\exp(irx)$ as $n \to \infty$.

This suggests a general way of constructing spectral estimates in situations where part of the data is missing. First an estimate of $R_X(r)$ is formed, by summing all the lagged products of the form $X(t)X(t+r)$ and dividing the sum by a suitable normalising factor. The resulting estimate of $R_X(r)$ is then substituted for $c_X(r)$ in (1.1.7) to give an estimate of $g_X(x)$. From the way in which this estimate is constructed, it is called a natural estimate.

In general, the distribution of a natural estimate is difficult to determine. It may be fairly easy to find its asymptotic mean, because of the way it is constructed. The asymptotic variance depends on the particular situation, and will be derived for some fairly general models.

An incomplete series is represented as follows. The amplitude modulating process, $\{Y(t)\}$, is defined by

$$Y(t) = \begin{cases} 1 & \text{if } X(t) \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

The amplitude modulated process $\{Z(t)\}$ is then given by $Z(t) = X(t)Y(t)$. Thus $\{Z(t)\}$ consists of the observed values of $\{X(t)\}$ with zeroes where values are missing.

It seems plausible to assume that often, when $\{Y(t)\}$

is a random series, it will be independent of $\{X(t)\}$. This will always be one of the basic assumptions in later chapters, which excludes discussion of any topics where this is not satisfied. Thus, for example, censored data, where observations are deleted if they lie outside certain limits, is not discussed.

CHAPTER 2. A STATIONARY RANDOM AMPLITUDE MODULATING PROCESS

2.1 Assumptions

The first case to be considered in any detail is one in which observations are missed in a fairly general random way. This is a natural extension of the work of Scheinok (1965). It is assumed that the amplitude modulating process is a weakly stationary process satisfying certain conditions. It is shown that the amplitude modulated process $\{Z(t)\}$ is also weakly stationary, and that it satisfies the conditions of Section 1.2, under which $g_Z(x)$, its spectral density function, may be estimated consistently. Two consistent estimates of $g_X(x)$, the spectral density function of the original process $\{X(t)\}$, are then constructed.

Suppose that $\{X(t)\}$ is a process satisfying (1.4) and (1.5), so that consistent estimation of $g_X(x)$ is possible from a complete set of observations. Suppose also that the mechanism which controls the missed observations gives rise to a random amplitude modulating process $\{Y(t)\}$ satisfying

$$Y(t) \text{ is independent of } X(s) ; \tag{2.1}$$

$$P\{Y(t) = 1\} = p = 1 - P\{Y(t) = 0\} , \quad \text{for some } p \text{ with } 0 < p < 1 ; \tag{2.2}$$

$$E\{Y(t)Y(t+r)\} = pu_r ; \tag{2.3}$$

$$E\{Y(t)Y(t+q)Y(t+r)Y(t+s)\} = p^2 v_{q,r,s} \qquad (2.4)$$

for $q,r,s,t = 0, \pm 1, \ldots$.

Some restrictions are made on these quantities, their form depends on whether or not the sequence $\{pu_r\}$ is assumed to be known, since a slightly different estimate of $g_X(x)$ is used in each case. If $\{pu_r\}$ is known, it is assumed that

$$u_r > 0 ; \qquad (2.5)$$

$$\sum_{q=-\infty}^{\infty} |v_{r,q,q+s} - u_r u_s| \leq V(|r| + |s| + 1) \qquad (2.6)$$

for $r,s = 0,\pm 1, \ldots$. If $\{pu_r\}$ is unknown, it is assumed that

$$pu_r > c \text{ for known } c \text{ satisfying } 0 < c < p^2; \qquad (2.7)$$

$$\sum_{q} |v_{r,q,q+s} - u_r u_s| \leq V \qquad (2.8)$$

for $r,s = 0,\pm 1, \ldots$.

The definitions (2.2) and (2.3) impose certain restrictions on p and $\{u_r\}$. From (2.2) it is clear that $0 \leq p \leq 1$. The extreme values are excluded in (2.2) as they correspond to degenerate cases, in which $Y(t)$ is identically zero or unity. From (2.3) it follows that $pu_r - p^2 = \mathrm{Cov}\{Y(t),Y(t+r)\}$, and so $\{pu_r - p^2\}$ is a positive definite sequence. This implies, for example, that

$$|pu_r - p^2| \leqslant pu_0 - p^2 = p(1-p)$$

or $2p - 1 \leqslant u_r \leqslant 1$. However, if $p \neq \frac{1}{2}$, the lower bound may be improved. For if $p < \frac{1}{2}$, $0 \leqslant u_r$ is better, while if $p > \frac{1}{2}$,

$$0 \leqslant E\left[\{1 - Y(t)\}\{1 - Y(t+r)\}\right] = 1 - 2p + pu_r$$

implies that $2p - 1 \leqslant pu_r$, which is better since $p < 1$. Incidentally, this shows that a sufficient condition for (2.5) to hold is $p > \frac{1}{2}$. Conditions (2.7) and (2.8) are stronger than (2.5) and (2.6). In a sense, this compensates for the lack of information about $\{pu_r\}$. Assumption (2.5) ensures that $R_X(r)$ can be estimated by a sum of lagged products, that is by an estimate of the form

$$k \sum X(t)X(t+r) ,$$

where k is a normalising factor. For if $u_s = 0$ for some s, it follows that $Y(t)Y(t+s) = 0$ with probability one. Thus if $X(t)$ is observed, $X(t+s)$ will certainly be missed, and no lagged products $X(t)X(t+s)$ will be available.

Assumption (2.6) implies that

$$\mathrm{Cov}\{Y(t_1)Y(t_1+r), Y(t_2)Y(t_2+s)\} \to 0$$

as $|t_1 - t_2| \to \infty$. If r and s are replaced by zero, (2.6) becomes

$$\sum_{q} |u_q - p| \leq pV < \infty , \tag{2.9}$$

which implies that $\mathrm{Cov}\{Y(t_1), Y(t_2)\} \to 0$ as $|t_1 - t_2| \to \infty$. Thus (2.6) ensures that the dependence between widely separated parts of the process $\{Y(t)\}$ is small in this sense. Expression (2.9) implies that $u_r \to p$ as $r \to \infty$. When $\{pu_r\}$ is unknown, that is (2.7) holds, this means that for some $\delta > 0$ and for each r,

$$pu_r \geq c + \delta . \tag{2.10}$$

Clearly $\{Y(t)\}$ is weakly stationary, with a positive mean. From (2.9), it follows that its auto-covariance sequence is summable, and hence it has an absolutely continuous spectral distribution function and a continuous spectral density function. Thus this model cannot contain any periodic components. Some models containing periodic components are discussed in later chapters.

The model used by Scheinok (1965) is a special case of the model described above. Scheinok assumed (2.1) and (2.2), and that $Y(t)$ is independent of $Y(s)$ for $t \neq s$. In the current notation, this means that $u_r = p$ for $r \neq 0$, and $v_{q,r,s} = p^2$ for $0 \neq q \neq r \neq s$. The summation conditions are easily shown to hold, as only a finite number of terms are nonzero in each case.

2.2 The Amplitude Modulated Process

Before proceeding to the estimation of $g_X(x)$, some properties of $\{Z(t)\}$, the amplitude modulated process, are derived, as they are useful later. First, note that $E\{Z(t)\} = 0$, and so

$$\begin{aligned} \mathrm{Cov}\{Z(t), Z(t+r)\} &= E\{Z(t)Z(t+r)\} \\ &= p u_r \, R_X(r) \\ &= R_Z(r) \end{aligned} \tag{2.11}$$

for each t . Also

$$R_Z = \sum_r |r R_Z(r)| \leq \sum_r |r R_X(r)| = R_X < \infty \tag{2.12}$$

from (1.4), whence $\{Z(t)\}$ also satisfies (1.4). The fourth cumulant function of $\{Z(t)\}$, $Q_Z(q,r,s)$, satisfies

$$\begin{aligned} Q_Z(q,r,s) &= p^2 \, Q_X(q,r,s) v_{q,r,s} \\ &\quad + R_X(q) R_X(r-s)(v_{q,r,s} - u_q \, u_{r-s}) \\ &\quad + R_X(r) R_X(s-q)(v_{q,r,s} - u_r \, u_{s-q}) \\ &\quad + R_X(s) R_X(q-r)(v_{q,r,s} - u_s \, u_{q-r}) \, , \end{aligned}$$

$q,r,s = 0, \pm 1, \ldots$. The fact that this is independent of t, together with (2.11) and $E\{Z(t)\} = 0$, shows that $\{Z(t)\}$ is weakly stationary.

On using (1.5) and (2.6), it may be shown that

$$Q_Z = \sum_{q,r,s=-\infty}^{\infty} Q_Z(q,r,s) \leq \infty \, . \tag{2.13}$$

Thus although the distribution of $Z(t)$ is distinctly non-Gaussian, as $P\{Z(t) = 0\} > 0$, the fourth cumulant function is still sufficiently small for (2.13), which is the same as (1.5), to hold.

It follows from (2.12) that $g_Z(x)$, the spectral density function of $\{Z(t)\}$, exists, and from (2.12) and (2.13), it may be estimated as described in Section 1.2. That is, if $\{w(y;x)\}$ is a sequence of weighting functions with the required properties, then

$$g_Z^{(1)}(x) = (2\pi)^{-1} \sum_{|r|<n} c_Z(r) w_r(x)$$

is a consistent estimate of $g_Z(x)$ where $c_Z(r)$ is defined as in (1.7), that is

$$c_Z(r) = n^{-1} \sum_{t=1}^{n-|r|} Z(t)Z(t+r) \ .$$

Also, from (1.15)

$$\mathrm{Var}\{g_Z^{(1)}(x)\} = n^{-1} W_n \, g_Z(x)^2 \, d(x) + o(n^{-1} W_n) .$$

Now $g_Z(x)$ can be expressed in terms of $g_X(x)$. For

$$R_Z(r) = pu_r \, R_X(r) = p^2 R_X(r) + p(u_r - p)R_X(r) \ ,$$

and thus if

$$a(x) = (2\pi)^{-1} p \sum_r (u_r - p)e^{irx} \ ,$$

where the sum converges by (2.9), then

$$g_Z(x) = (2\pi)^{-1} \sum_r R_Z(r) e^{irx}$$

$$= p^2 g_X(x) + \int_{-\pi}^{\pi} a(x-y) g_X(y) dy \ . \qquad (2.14)$$

The integral in (2.14) is positive since $a(x)$ is the spectral density function of $Y(t)$. Hence

$$g_Z(x) \geqslant p^2 g_X(x) \ . \qquad (2.15)$$

Note also that

$$g_Z(x) \leqslant p^2 \sup_x g_X(x) + \sup_x g_X(x) \int_{-\pi}^{\pi} a(x-y) dy$$

$$= p \sup_x g_X(x)$$

and similarly $p \inf_x g_X(x) \leqslant g_Z(x)$. Thus $g_Z(x)$ is smoother than $g_X(x)$ in the sense that

$$\sup_x g_X(x) / \inf_x g_X(x) \geqslant \sup_x g_Z(x) / \inf_x g_Z(x) \ .$$

Equation (2.14) can be solved for $g_X(x)$ in terms of $g_Z(x)$. Let

$$b_r = p^{-2} - (pu_r)^{-1} = (u_r - p)/(p^2 u_r) . \qquad (2.16)$$

Then b_r is finite and

$$\sum_r |b_r| < \infty \qquad (2.17)$$

by (2.5) and (2.9). Let

$$b(x) = (2\pi)^{-1} \sum_r b_r e^{irx} \ .$$

Then

$$g_X(x) = g_Z(x)/p^2 - \int_{-\pi}^{\pi} b(x-y)\, g_Z(y)\, dy \ .$$

The formulae for $a(x)$ and $b(x)$ are particularly simple when $Y(s)$ and $Y(t)$ are independent for $s \neq t$. For then $u_r = p$ for $r \neq 0$, and hence

$$a(x) = (2\pi)^{-1}\, p(1-p) \ ,$$

$$b(x) = (2\pi)^{-1}\, (1-p)/p^2$$

are both constants.

2.3 An Estimate of $g_X(x)$

In this section, an estimate of $g_X(x)$ is constructed on the assumption that the sequence $\{pu_r\}$ is known. This gives rise to the simplest estimate. In the next section this assumption is dropped.

Following the procedure outlined in Section 1.3, the sum of all available lagged products $X(t)X(t+r)$ is formed. This sum is $nc_Z(r)$, where

$$c_Z(r) = n^{-1} \sum_{t=1}^{n-|r|} Z(t)Z(t+|r|) \ ,$$

as in (1.7).

Since $\{X(t)\}$ and $\{Y(t)\}$ are independent, let E_X and E_Y denote the expectation operators for the two sequences. Later we shall use a similar notation for variances

and covariances. Then

$$E_Y\{c_Z(r)\} = pu_r\, c_X(r) \ .$$

Now $c_X(r)$ is the usual estimate of $R_X(r)$ from the complete data. Thus $(pu_r)^{-1}\, c_Z(r)$ is a suitable estimate of $R_X(r)$ from the incomplete data $\{Z(t)\}$. If this estimate is substituted for $c_X(r)$ in (1.17), the estimate

$$g_X^{(2)}(x) = (2\pi)^{-1} \sum_{|r|<n} (pu_r)^{-1}\, c_Z(r) w_r(x) \qquad (2.18)$$

is obtained. Using (2.16), this may be written

$$g_X^{(2)}(x) = g_Z^{(1)}(x)/p^2 - (2\pi)^{-1} \sum_{|r|<n} b_r\, c_Z(r) w_r(x)$$

$$= g_Z^{(1)}(x)/p^2 - T(x) \ , \qquad (2.19)$$

say.

Expression (2.18) is the more convenient form for computation, but (2.19) enables the asymptotic variance of $g_X^{(2)}(x)$ to be found quite easily. Similar methods are used later, and so this proof is given in some detail. Now from (2.19),

$$\mathrm{Var}\{g_X^{(2)}(x)\} = \mathrm{Var}\{g_Z^{(1)}(x)\}/p^4 - 2\mathrm{Cov}\{g_Z^{(1)}(x), T(x)\}/p^2$$

$$+ \mathrm{Var}\{T(x)\} \ .$$

The second term on the right hand side is a multiple of

$$\mathrm{Cov}\{g_Z^{(1)}(x), T(x)\} = (2\pi n)^{-2}\, \mathrm{Cov}\Big\{ \sum_{q,r=1}^{n} Z(q)Z(r)w_{q-r}(x), \sum_{s,t=1}^{n} Z(s)Z(t)b_{s-t}\, w_{s-t}(x) \Big\}$$

$$= (2\pi n)^{-2} \sum_{q,r,s,t=1}^{n} \Big[w_{q-r}(x)\overline{w_{s-t}(x)}\, b_{s-t} \cdot \mathrm{Cov}\{Z(q)Z(r), Z(s)Z(t)\} \Big] ,$$

where the bar denotes complex conjugation. Now, from the definition of $Q_Z(q,r,s)$,

$$\mathrm{Cov}\{Z(q)Z(r), Z(s)Z(t)\} = Q_Z(r-q, s-q, t-q) + R_Z(s-q)R_Z(t-r) + R_Z(s-r)R_Z(t-q) ,$$

and $\mathrm{Cov}\{g_Z^{(1)}(x), T(x)\}$ similarly splits up into three sums, say T_1 , T_2 , and T_3 . Then

$$|T_1| = \Big| (2\pi n)^{-2} \sum_{q,r,s,t=1}^{n} \Big\{ w_{q-r}(x)\overline{w_{s-t}(x)}\, b_{s-t} \cdot Q_Z(r-q, s-q, t-q) \Big\} \Big|$$

$$\leq M^2 (2\pi n)^{-2} \sum_{q,r,s,t=1}^{n} \big| b_{s-t}\, Q_Z(r-q, s-q, t-q) \big|$$

$$\leq M^2 (2\pi n)^{-2} \sum_{q=1}^{n} \sum_{r,s,t=-\infty}^{\infty} \big| b_{s-t}\, Q_Z(r-q, s-q, t-q) \big| ,$$

where the first inequality uses (1.13).

Now $\{b_r\}$ is a bounded sequence, and so the triple sum converges by (2.13) to a value which is clearly independent of q and n, say S. Thus

$$|T_1| \leq M^2(2\pi n)^{-2} \sum_{q=1}^{n} S$$
$$= \{M^2(2\pi)^{-2} S\} n^{-1}$$

and so $T_1 = O(n^{-1})$. Similarly

$$|T_2| \leq M^2(2\pi n)^{-2} \sum_{q,r,s,t=1}^{n} |b_{s-t} R_Z(s-q)R_Z(t-r)|$$

Put $j = s-t$, $k = s-q$, $\ell = t-r$, $m = q$. Then

$$|T_2| \leq M^2(2\pi n)^{-2} \sum_{m=1}^{n} \sum_{j,k,\ell=-\infty}^{\infty} |b_j R_Z(k)R_Z(\ell)|$$

As before, the triple sum converges to a value independent of m or n, because of (2.12) and (2.17), and so $T_2 = O(n^{-1})$. The term T_3 is handled in the same way as T_2, and may also be shown to be $O(n^{-1})$. These three results imply that $\mathrm{Cov}\{g_Z^{(1)}(x), T(x)\} = O(n^{-1})$. In a similar way it may be shown that $\mathrm{Var}\{T(x)\} = O(n^{-1})$, so that finally

$$\mathrm{Var}\{g_X^{(2)}(x)\} = \mathrm{Var}\{g_Z^{(1)}(x)\}/p^4 + O(n^{-1}) . \qquad (2.20)$$

It has already been pointed out that $\mathrm{Var}\{g_Z^{(1)}(x)\}$ decreases more slowly than n^{-1}, and hence is the dominant term. Thus, from (1.15)

$$\mathrm{Var}\{g_X^{(2)}(x)\} = p^{-4}n^{-1}W_n g_Z(x)^2 d(x) + o(n^{-1}W_n) . \quad (2.21)$$

In a similar way it may be shown that

$$\mathrm{Cov}\{g_X^{(2)}(x_1), g_X^{(2)}(x_2)\} = o(n^{-1} W_n)$$

if $x_1 \neq \pm x_2$.

From the construction of $g_X^{(2)}(x)$, it has the same mean as $g_X^{(1)}(x)$, and hence is asymptotically unbiased. This, with (2.21), shows that if $n^{-1} W_n \to 0$ as $n \to \infty$, then $g_X^{(2)}(x)$ is a consistent estimate of $g_X(x)$. The only unknown quantity in its asymptotic variance is $g_X(x)$, which may also be estimated consistently.

It is worth recalling here that the asymptotic variance of $g_X^{(1)}(x)$, the estimate based on complete data, is

$$n^{-1} W_n g_X(x)^2 d(x) \leqslant n^{-1} W_n p^{-4} g_Z(x)^2 d(x)$$

by (2.15). Thus the asymptotic variance is always increased by the absence of data. An upper bound for the increase is provided by the inequality

$$g_Z(x)^2 \leqslant p^2 \sup g_X(x)^2 .$$

Note also that the structure of $\{Y(t)\}$ affects the

asymptotic variance of $g_X^{(2)}(x)$ only through p and $g_Z(x)$; that is through p and a(x) . In other words, only the mean and spectral density function of $\{Y(t)\}$ affect the result. This seems to be essentially due to the fact that $\{Y(t)\}$ has an absolutely continuous spectral distribution function. In some models to be discussed later involving periodic effects, other properties of the amplitude modulating process affect the asymptotic variance of the estimate.

2.4 A More Useful Estimate

In the previous section, $g_X^{(2)}(x)$ was shown to be a consistent estimate of $g_X(x)$. However, in the construction, it was assumed that $\{pu_r\}$ was known, which is unlikely to be the case in general. In this section a consistent estimate of $g_X(x)$ is constructed without that assumption.

If

$$C_Y^*(r) = c_Y(r)/(1 - n^{-1}|r|)$$

$$= (n - |r|)^{-1} \sum_{t=1}^{n-|r|} Y(t)Y(t+|r|) \quad , \qquad (2.22)$$

then $E\{C_Y^*(r)\} = pu_r$. Thus it would appear that a possible procedure would be to replace pu_r in (2.18) by $C_Y^*(r)$. Now in general $P\{C_Y^*(r) = 0\} > 0$. As $C_Y^*(r)$ is a sum of nonnegative terms, $C_Y^*(r) = 0$ would imply that each term is zero, and hence that $c_Z(r) = 0$. Thus the ratio $c_Z(r)/C_Y^*(r)$ would be undefined, and could be assigned some neutral value,

such as zero. However, it has not been possible to derive the asymptotic variances of the resulting estimate of $g_X(x)$.

The difficulty is caused by small positive values of $C_Y^*(r)$. The estimate of pu_r is modified to exclude these as follows. Let c be as specified in (2.7), and let

$$C_Y(r) = \begin{cases} c & \text{if } C_Y^*(r) \leq c \ , \\ C_Y^*(r) & \text{otherwise.} \end{cases} \qquad (2.23)$$

The estimate $C_Y(r)$ is now substituted for pu_r in (2.18), to give an estimate $g_X^{(3)}(x)$ of $g_X(x)$, where

$$g_X^{(3)}(x) = (2\pi)^{-1} \sum_{|r|<n} C_Y(r)^{-1} c_Z(r) w_r(x) \ . \qquad (2.24)$$

It is shown below that $g_X^{(3)}(x)$ is a consistent estimate of $g_X(x)$.

It was a theoretical difficulty which led to the use of a modified estimate of pu_r . Whether or not this reflects any practical difficulty in the use of the unmodified estimate has not been investigated. If in an application the unmodified estimate did give trouble, the choice of c might prove to be critical. This also has not been investigated. Neither problem should cause much trouble if $1 - p$ were small.

This estimate of $g_X(x)$ is intuitively appealing. For $C_Y^*(r)$ is the ratio of the number of nonzero terms in $c_Z(r)$

in the incomplete sample to the number of terms in $c_X(r)$ in a full sample, and if $1 - p$ is small, $C_Y(r)$ will usually be the same value. Indeed, it is to be expected that the estimate will be better than $g_X^{(2)}(x)$, as it uses information about missing observations from the sample, rather than ensemble properties of $\{Y(t)\}$. It is shown below that $g_X^{(3)}(x)$ has the same asymptotic mean and variance as $g_X^{(2)}(x)$. This does not, of course, exclude the possibility that $g_X^{(3)}(x)$ would be better in small sample cases.

Let

$$D = g_X^{(2)}(x) - g_X^{(3)}(x)$$

$$= (2\pi)^{-1} \sum_{|r|<n} c_Z(r)\{(pu_r)^{-1} - C_Y(r)^{-1}\}w_r(x). \quad (2.25)$$

We now show that $E(D)$ and $\mathrm{Var}(D)$ are both $O(n^{-1})$. It then follows that

$$E\{g_X^{(3)}(x)\} = E\{g_X^{(2)}(x)\} + O(n^{-1}),$$

$$\mathrm{Var}\{g_X^{(3)}(x)\} = \mathrm{Var}\{g_X^{(2)}(x)\} + o(n^{-1}W_n),$$

so that $g_X^{(3)}(x)$ has the same asymptotic mean and variance as $g_X^{(2)}(x)$.

As before, a subscript X or Y attached to the expectation and variance operators indicates that the operation is only applied to the corresponding process. Where an expression T involves only one process, say $\{X(t)\}$,

it is clear that $E(T) = E_X(T)$ etc., and the subscript will usually be omitted.

The following results are needed. First, for $r \geqslant 0$,

$$\begin{aligned}
\mathrm{Var}\{C_Y^*(r)\} &= \mathrm{Var}\Big\{(n-r)^{-1}\sum_{t=1}^{n-r} Y(t)Y(t+r)\Big\} \\
&= (n-r)^{-2}\sum_{s,t=1}^{n-r} \mathrm{Cov}\{Y(s)Y(s+r), Y(t)Y(t+r)\} \\
&= p^2(n-r)^{-2}\sum_{s,t=1}^{n-r}(v_{r,t-s,t-s+r} - u_r^2) \\
&\leqslant p^2(n-r)^{-2}\sum_{s=1}^{n-r}\sum_{t=-\infty}^{\infty}|v_{r,t-s,t-s+r} - u_r^2| \\
&\leqslant p^2(n-r)^{-2}\sum_{s=1}^{n-r} V \\
&= p^2 V/(n-r) ,
\end{aligned}$$

where the second inequality uses (2.8). Thus, for any r,

$$\mathrm{Var}\{C_Y^*(r)\} \leqslant p^2 V/(n-|r|) . \qquad (2.26)$$

Also

$$\begin{aligned}
E\Big[\{(pu_r)^{-1} - C_Y(r)^{-1}\}^2\Big] &= (pu_r)^{-2} E\Big[\{C_Y(r) - pu_r\}^2/C_Y(r)^2\Big] \\
&\leqslant c^{-4} E\Big[\{C_Y(r) - pu_r\}^2\Big] \\
&\leqslant c^{-4} E\Big[\{C_Y^*(r) - pu_r\}^2\Big] \\
&= c^{-4}\,\mathrm{Var}\{C_Y^*(r)\}
\end{aligned}$$

by (2.7) and (2.23). Hence

$$E\left[\left\{(pu_r)^{-1} - C_Y(r)^{-1}\right\}^2\right] \leq p^2\, Vc^{-4}\, (n - |r|)^{-1}, \quad (2.27)$$

and consequently

$$E\left\{\left|(pu_r)^{-1} - C_Y(r)^{-1}\right|\left|(pu_s)^{-1} - C_Y(s)^{-1}\right|\right\}$$
$$\leq p^2\, Vc^{-4}\, (n - |r|)^{-\frac{1}{2}}\, (n - |s|)^{-\frac{1}{2}}, \quad (2.28)$$

using (2.27) and the Cauchy-Schwarz inequality.

We now show that $E(D) = O(n^{-1})$. For

$$E_X(D) = (2\pi)^{-1} \sum_{|r|<n} E_X\{c_Z(r)\}\left\{(pu_r)^{-1} - C_Y(r)^{-1}\right\} w_r(x)$$
$$= (2\pi n)^{-1} \sum_{|r|<n} \left[(n - |r|) R_X(r) C_Y^*(r) \cdot \left\{(pu_r)^{-1} - C_Y(r)^{-1}\right\} w_r(x)\right], \quad (2.29)$$

and hence

$$E(D) = E_Y\{E_X(D)\}$$
$$= (2\pi n)^{-1} \sum_{|r|<n} \left\{(n - |r|) R_X(r) w_r(x) \cdot \left[1 - E\{C_Y^*(r)/C_Y(r)\}\right]\right\}. \quad (2.30)$$

But

$$0 \leq C_Y^*(r)/C_Y(r) = \begin{cases} 1 & \text{if } C_Y^*(r) > c, \\ C_Y^*(r)/c \leq 1 & \text{otherwise,} \end{cases}$$

and so

$$1 \geq E\{C_Y^*(r)/C_Y(r)\} \geq 1 - P\{C_Y^*(r) \leq c\}.$$

Using Chebychev's inequality and (2.26), we find that

$$0 \leq 1 - E\{C_Y^*(r)/C_Y(r)\} \leq p^2 V/\{(n-|r|)(pu_r - c)^2\}$$

or, using (2.10),

$$|1 - E\{C_Y^*(r)/C_Y(r)\}| \leq p^2 V/\{(n-|r|)\delta^2\} .$$

Thus from (1.13), (2.12) and (2.30),

$$|E(D)| \leq (2\pi n)^{-1} \sum_{|r|<n} |R_X(r)| \, p^2 VM/\delta^2$$

$$\leq p^2 VM\{R_X + R_X(0)\}/(2\,n\delta^2)$$

$$= O(n^{-1}) .$$

Hence we have shown that $E(D) = O(n^{-1})$. This could have been proved using (2.6), the weaker version of (2.8). The stronger condition is required in the next part, where we examine $Var(D)$.

We use the relationship

$$Var(D) = Var_Y\{E_X(D)\} + E_Y\{Var_X(D)\} , \qquad (2.31)$$

and examine first the term $Var_Y\{E_X(D)\}$. Now $0 \leq C_Y^*(r) \leq 1$, and so from (1.13) and (2.29),

$$|E_X(D)| \leq M(2\pi n)^{-1} \sum_{|r|<n} |R_X(r)| \, (n-|r|) \left| (pu_r)^{-1} - C_Y(r)^{-1} \right| .$$

Thus

$$Var_Y\{E_X(D)\} \leq E_Y\{E_X(D)^2\}$$

$$\leq M^2(2\pi n)^{-2} \sum_{|r|,\,|s|<n} \Big[|R_X(r)R_X(s)| \,(n-|r|)(n-|s|) \cdot E\big\{ |(pu_r)^{-1} - C_Y(r)^{-1}|\,|(pu_s)^{-1} - C_Y(s)^{-1}| \big\} \Big]$$

$$\leq M^2(2\pi n)^{-2} \sum_{|r|,\,|s|<n} \Big[|R_X(r)R_X(s)| \,(n-|r|)(n-|s|) \cdot p^2 V / \{c^4 (n-|r|)^{\frac{1}{2}}(n-|s|)^{\frac{1}{2}}\} \Big]$$

using (2.28), whence, from (2.12),

$$\mathrm{Var}_Y\{E_X(D)\} \leq M^2 p^2 V n^{-1} \sum_{|r|,\,|s|<n} |R_X(r)R_X(s)|/c^4$$

$$= O(n^{-1}) \,. \qquad (2.32)$$

To deal with the second term of (2.31), we rewrite D as

$$D = (2\pi n)^{-1} \sum_{s,\,t=1}^{n} Z(s)Z(t)\{(pu_{s-t})^{-1} - C_Y(s-t)^{-1}\} w_{s-t}(x),$$

from which

$$\mathrm{Var}_X(D) = (2\pi n)^{-2} \sum_{q,r,s,t=1}^{n} \Big[Y(q)Y(r)Y(s)Y(t) \cdot \{(pu_{q-r})^{-1} - C_Y(q-r)^{-1}\} w_{q-r}(x) \cdot \{(pu_{s-t})^{-1} - C_Y(s-t)^{-1}\} \overline{w_{s-t}(x)} \cdot \mathrm{Cov}\{X(q)X(r), X(s)X(t)\} \Big] \,.$$

Now

$$\mathrm{Cov}\{X(q)X(r),X(s)X(t)\} = Q_X(r-q,s-q,t-q) + R_X(s-q)R_X(t-r) + R_X(s-r)R_X(t-q) ,$$

and when this is substituted in the previous expression, $\mathrm{Var}_X(D)$ splits up into three corresponding sums, T_1 , T_2 and T_3 , say. Then, using (1.13) and the fact that $|Y(t)| \leq 1$ for each t ,

$$|T_1| \leq M^2(2\pi n)^{-2} \sum_{q,r,s,t=1}^{n} \left\{ |Q_X(r-q,s-q,t-q)| \cdot \left| (pu_{q-r})^{-1} - C_Y(q-r)^{-1} \right| \left| (pu_{s-t})^{-1} - C_Y(s-t)^{-1} \right| \right\} .$$

From (2.28),

$$E_Y(|T_1|) \leq M^2(2\pi n)^{-2} \sum_{q,r,s,t=1}^{n} \left\{ |Q_X(r-q,s-q,t-q)| \cdot p^2 Vc^{-4} (n-|q-r|)^{-\frac{1}{2}} (n-|s-t|)^{-\frac{1}{2}} \right\}$$

$$\leq M^2(2\pi n)^{-2} p^2 Vc^{-4} \sum_{q,r,s,t=1}^{n} |Q_X(r-q,s-q,t-q)| .$$

As in the previous section, it may be shown that the quadruple sum is $O(n)$, whence

$$|E_Y(T_1)| \leq E_Y(|T_1|) = O(n^{-1}) . \qquad (2.33)$$

Also

$$T_2 = (2\pi n)^{-2} \sum_{q,r,s,t=1}^{n} \Big[Y(q)Y(r)Y(s)Y(t)\ R_X(q-s)R_X(r-t) \cdot \{(pu_{q-r})^{-1} - C_Y(q-r)^{-1}\} w_{q-r}(x) \cdot \{(pu_{s-t})^{-1} - C_Y(s-t)^{-1}\} \overline{w_{s-t}(x)} \Big].$$

We change the summation variables to $j = q - s$, $k = r - t$, $\ell = s - t$, $m = t$, to give

$$T_2 = (2\pi n)^{-2} \sum_{|j|,|k|<n} \Big[R_X(j)R_X(k) \sum_{\ell} \{(pu_{\ell})^{-1} - C_Y(\ell)^{-1}\} w_{\ell}(x) \cdot \{(pu_{j-k+\ell})^{-1} - C_Y(j-k+\ell)^{-1}\} \overline{w_{j-k+\ell}(x)} \cdot \sum_{m} Y(m)Y(m+\ell)Y(m+k)Y(m+\ell+j) \Big],$$

where the range of summation over ℓ and m is such that the argument of each Y lies in the set $\{1, 2, \ldots, n\}$. Thus, since $Y(t)^2 = Y(t)$ with unit probability, and using the Cauchy-Schwarz inequality,

$$\begin{aligned} &\sum_{m} Y(m)Y(m+\ell)Y(m+k)Y(m+\ell+j) \\ &\leq \Big\{ \sum_{m} Y(m)Y(m+\ell) \sum_{m} Y(m+k)Y(m+\ell+j) \Big\}^{\frac{1}{2}} \\ &\leqslant (n - |\ell|)^{\frac{1}{2}} (n - |j-k+\ell|)^{\frac{1}{2}}, \end{aligned}$$

with unit probability, and hence from (1.13),

$$|E(T_2)| \leq E(|T_2|)$$

$$\leq M^2(2\pi n)^{-2} \sum_{|j|,|k|<n} \left(|R_X(j)R_X(k)| \sum_{\ell} \left[(n-|\ell|)^{\frac{1}{2}} (n-|j-k+\ell|)^{\frac{1}{2}} \right.\right.$$

$$\left.\left. \cdot E_Y\left\{ \left| (pu_\ell)^{-1} - C_Y(\ell)^{-1} \right| \left| (pu_{j-k+\ell})^{-1} - C_Y(j-k+\ell)^{-1} \right| \right\} \right] \right)$$

$$\leq M^2(2\pi n)^{-2} \sum_{|j|,|k|<n} |R_X(j)R_X(k)| \sum_{\ell} p^2 V\, c^{-4}$$

from (2.28). Now the summation over $\ell = s - t$ contains less than $2n$ terms, and thus the summation over j and k is $O(n)$. Hence finally

$$E(T_2) = O(n^{-1}) .$$

The last term, T_3, may in a similar way be shown to be $O(n^{-1})$. Combining this with (2.33), we see that $E_Y\{\mathrm{Var}_X(D)\} = O(n^{-1})$, which with (2.31) and (2.32) implies that $\mathrm{Var}(D) = O(n^{-1})$, which was to be proved.

Thus it has been proved that $g_X^{(3)}(x)$ has the same asymptotic mean and variance as $g_X^{(2)}(x)$. To be exact,

$$E\{g_X^{(3)}(x)\} = E\{g_X^{(2)}(x)\} + O(n^{-1}), \tag{2.34}$$

$$\mathrm{Var}\{g_X^{(3)}(x)\} = \mathrm{Var}\{g_X^{(2)}(x)\} + o(n^{-1}W_n)$$

$$= n^{-1}W_n\, g_Z(x)^2 d(x) + o(n^{-1}W_n) \tag{2.35}$$

from (2.21). In a similar way it may be shown that

$$\mathrm{Cov}\{g_X^{(3)}(x_1), g_X^{(3)}(x_2)\} = o(n^{-1}W_n)$$

if $x_1 \neq x_2$.

2.5 Two Specific Cases

In this section, two models are discussed which describe missing data situations. Conditions are found under which they satisfy the assumptions of Section 2.1. Each model has one feature in common with the independent case considered by Scheinok (1965); he assumed that an observation is made with probability p independently of the other observations. This gives rise to strings of observations of mean length $p/(1-p)$, separated by gaps of mean length $(1-p)/p$, strings and gaps being independently geometrically distributed.

In the first model, the intervals between successive observations are independently distributed with a common arbitrary distribution. Thus observations occur in a renewal process, and strings of observations are geometrically distributed. This model describes for instance a situation where some recording apparatus fails with constant probability at each time point when it is functioning, and repair times are independent. This is referred to as the renewal model.

Suppose that the intervals between successive observations have probability distribution $\{f_r : r = 1,2,\ldots\}$, probability generating function $F(z) = \sum f_r z^r$, and finite mean $F'(1) = p^{-1}$. To satisfy (2.2), $f_1 < 1$. It may be verified that if the interval to the first observation has the probability generating function $p\{1-F(z)\}/(1-z)$, then the resulting amplitude modulating process $\{Y(t)\}$ is stationary.

To find the properties of $\{Y(t)\}$, define a process $\{Y^{0}(t)\}$ to have the same structure except that the initial condition is that $Y^{0}(0) = 1$. Then the event $\mathcal{E}$: "$Y^{0} = 1$" is a persistent recurrent event, as defined by Feller (1968, p.308)

Now from (2.3),

$$\begin{aligned} u_r &= E\{Y(t)Y(t+r)\}/p \\ &= P\{Y(t+r) = 1 \,|\, Y(t) = 1\} \\ &= P\{Y^{0}(r) = 1\} \end{aligned}$$

and from the construction of $\{Y^{0}(t)\}$,

$$f_r = P\{Y^{0}(t) = 0, t = 1, \ldots, r-1; Y^{0}(r) = 1\} .$$

Hence the definitions of $\{u_r : r = 0,1, \ldots\}$ and $\{f_r : r = 1,2, \ldots\}$ coincide with those of Feller. Thus, for $r \geqslant 0$,

$$u_r = \sum_{s=1}^{r} f_s \, u_{r-s} . \qquad (2.36)$$

This implies that $u_r \geqslant f_1 \, u_{r-1}$ with equality for $r = 1$; therefore (2.5) is satisfied if and only if $f_1 > 0$.

When $f_1 > 0$, $\mathcal{E}$ is not periodic, and so $u_r \to p$ as $r \to \infty$ (Feller, 1968, p.335). However, a stronger condition is required to hold, namely (2.9). We show now that in this case, (2.9) implies (2.6), and, with an extra condition, also implies (2.8).

Suppose that $0 \leq q \leq r \leq s$. Then

$$p^2 v_{q,r,s} = P\{Y(t) = 1, Y(t+q) = 1, Y(t+r) = 1, Y(t+s) = 1\} .$$
$$= u_{s-r} u_{r-q} u_q p .$$

Hence, if $0 \leq r \leq s$,

$$p \sum_{q=-\infty}^{\infty} |v_{r,q,q+s} - u_r u_s|$$

$$= p\left\{\sum_{-\infty}^{-r} + \sum_{-r+1}^{0} + \sum_{1}^{s-r} + \sum_{s-r+1}^{s} + \sum_{s+1}^{\infty}\right\} |v_{r,q,q+s} - u_r u_s|$$

$$= \sum_{-\infty}^{-r} |u_{q+r} u_r u_s - p u_r u_s| + \sum_{-r+1}^{0} |u_q u_{r+q} u_{r+q-s} - p u_r u_s|$$

$$+ \sum_{1}^{s-r} |u_r u_q u_{r+q-s} - p u_r u_s| + \sum_{s-r+1}^{s} |u_q u_{q-s} u_{r+q-s} - p u_r u_s|$$

$$+ \sum_{s+1}^{\infty} |u_{q-s} u_r u_s - p u_r u_s|$$

$$\leq 3pV + 2(r|u_r - p| + s|u_s - p|).$$

We now use the symmetries of u_r and $v_{q,r,s}$ to state the general result that for any r and s,

$$p \sum_{q} |v_{r,q,q+s} - u_r u_s| \leq 3pV + 2\{|r(u_r - p)| + |s(u_s - p)|\} . \tag{2.37}$$

Thus (2.6) holds, because $0 \leq u_r \leq 1$, and we have shown that (2.2), (2.5) and (2.6) hold if and only if $0 < f_i < 1$ and (2.9) holds. Hence these are the conditions needed when $\{pu_r\}$ is known. One condition which implies (2.9) is discussed later in this section.

When $\{pu_r\}$ is unknown, the extra conditions (2.7) and (2.8) must be satisfied. From (2.37), (2.8) holds if (2.9) holds and $|r(u_r - p)| \leq U < \infty$ for each r. Assumption (2.7) requires that a positive lower bound for pu_r be known. It has been shown that $pu_r \geq 2p - 1$. Hence if $p > \frac{1}{2}$ is a known lower bound for p, any c satisfying $0 \leq c < 2p - 1$ will suffice.

Another possible way to construct c is as follows. Bloomfield (1970) shows that

$$pu_r \geq pu_1 \exp\{(1 - u_1)^{-1}(p^{-1} - 2 + u_1) \log u_1\} . \quad (2.37)$$

Thus if positive lower bounds for p and u_1 are known, a positive value of c can be constructed to satisfy (2.7).

One condition which ensures both that (2.9) holds and that $|r(u_r - p)|$ is bounded is that $f_r = O(\rho^{-r})$ as $r \to \infty$, for some $\rho > 1$. This would be satisfied, for instance, if the intervals between observations were essentially bounded. Although this is unnecessarily strong from a mathematical point of view, it is a mild restriction practically.

Suppose that $f_r = O(\rho^{-r})$ as $r \to \infty$ for some $\rho > 1$. Then $F(z)$ is analytic in $|z| < \rho$.

Let

$$U(z) = \sum_{r=0}^{\infty} u_r z^r .$$

Feller (1968, p.311) shows that $U(z) = \{1 - F(z)\}^{-1}$, and hence it may be shown that $U(z)$ has no poles in $|z| \leq 1$ apart from a simple pole at $z = 1$. Thus $A(z) = U(z) - p/(1 - z)$ is analytic in $|z| < \rho^*$ for some $\rho^* > 1$. But

$$A(z) = \sum_{r=0}^{\infty} (u_r - p)\, z^r,$$

and hence

$$\sum_{r=0}^{\infty} |u_r - p| < \infty,$$

$$\sum_{r=0}^{\infty} |r(u_r - p)| < \infty.$$

Thus (2.9) holds, and $|r(u_r - p)|$ is bounded.

The second model discussed in this section is closely related to the first. Suppose that the intervals between successive missed observations are independently distributed with a common arbitrary distribution. Then missed observations occur in a renewal process, and the gaps between observations are geometrically distributed. This would arise, for instance, if a recording apparatus failed after a time with some general distribution independently of the past, and repair times were geometrically distributed.

The relationship with the renewal model is clear. Symbols associated with the second model will be distinguished by an asterisk. Then the process $\{1 - Y^*(t)\}$ has the same structure as that of the process $\{Y(t)\}$ in the renewal model. Because of this similarity, the discussion of this

is brief, and consists only of a statement of conditions under which the various assumptions of Section 2.1 are satisfied.

Suppose that the intervals between successive missed observations have the distribution $\{f_r^*\}$. To satisfy (2.2), this distribution must have $f_1^* < 1$. It may also be shown that this is sufficient for (2.5) to hold.

As in the renewal model, (2.6) and (2.9) are equivalent, and (2.6) and (2.8) are both satisfied if $f_r^* = O(\rho^{-r})$ as $r \to \infty$, for some $\rho > 1$. If $p^* > \frac{1}{2}$ is a known lower bound for p^*, then any c satisfying $0 < c < 2p^* - 1$ will suffice in (2.7). If upper and lower bounds for p^* are known, and also a lower bound for u_1^*, another way to find such a c is through (2.37) and

$$p^* u_r^* = 1 - 2p + pu_r ,$$

where p and $\{u_r\}$ are the quantities associated with the process $\{1 - Y^*(t)\}$.

2.6 Nonstationary Deletion of Data

The effect of a particular type of nonstationarity in the deletion of data is discussed. Since this is contrary to one of the basic assumptions of Section 2.1, other effects of deviations from those assumptions are also discussed. The problems which arise when $\{X(t)\}$ does not satisfy the assumptions of Section 1.2 are no different from those in

the usual theory of spectral estimation; see for example Jenkins (1961) and Jenkins & Watts (1968, p.151).

The assumptions about the amplitude modulating process $\{Y(t)\}$ are of three types. The first is that it is independent of $\{X(t)\}$, the process under observation. This is assumed in order that certain product moments should factorise, which is the basis of the methods used. Thus any deviation from this assumption which seriously affected the factorisation would invalidate the results given earlier in this chapter. As has already been stated, this excludes topics such as censoring.

The third type of assumption, which for convenience is considered next, concerns the sum of an infinite series. Such assumptions are, of course, difficult to assess in the light of a finite amount of data. In the two examples considered in the previous section, the required assumptions were implied by a plausible condition. In the models involving periodic effects introduced in later chapters, it may be verified that these assumptions are not satisfied. This is also implied by the fact that different results are obtained. Thus it appears that these assumptions, possible weakened slightly, are essential to obtain the results.

The second type of assumption concerns the stationarity of the lower moments of $\{Y(t)\}$. One simple example of non-stationarity arises when part of the data is complete and

part is incomplete due to the action of a random amplitude modulating process. To obtain asymptotic results, it is assumed that an almost constant proportion of the data is complete.

To be explicit, suppose that $\{X(t)\}$ satisfies the conditions of Section 1.2, that $\{Y(t)\}$ satisfies those of Section 2.1 and that $\{A_n : n = 1,2, \ldots\}$ is a sequence of integers such that

$$1 \leq A_n \leq n ,$$
$$A_n/n = A + O(n^{-1}), \quad 0 < A < 1 . \tag{2.38}$$

Let

$$Y^{(n)}(t) = \begin{cases} 1 & 0 \leq t \leq A_n , \\ Y(t) & A_n < t \leq n , \end{cases}$$

and let $Z^{(n)}(t) = X(t)Y^{(n)}(t)$. The data available for analysis are $\{Z^{(n)}(t) : t = 1, \ldots , n\}$, and the proportion of the record which is complete is approximately A .

The amplitude modulating process $\{Y^{(n)}(t)\}$ is non-stationary, but possesses certain average properties. For instance,

$$n^{-1} \sum_{t=1}^{n} E\{Y^{(n)}(t)\} = A + (1 - A)p + O(n^{-1}) ,$$
$$n^{-1} \sum_{t=1}^{n-r} E\{Y^{(n)}(t)Y^{(n)}(t+r)\} = A + (1 - A)pu_r + O(n^{-1}r) \tag{2.39}$$

for $r \geq 0$.

Suppose that these data were analysed as if the amplitude modulating process were stationary. This would involve the use of the estimate $g_X^{(3)}(x)$ defined in (2.24). In order to simplify the discussion, it is assumed that the estimate $g_X^{(2)}(x)$ is used instead, with pu_r replaced by the average value (2.39). The estimate becomes

$$g_X^{(4)}(x) = (2\pi n)^{-1} \sum_{s,t=1}^{n} \left[Z^{(n)}(s) Z^{(n)}(t) w_{s-t}(x) \cdot \{A + (1 - A) pu_{s-t}\}^{-1} \right] . \tag{2.40}$$

The asymptotic variances derived below would be the same if an estimate based on $g_X^{(3)}(x)$ were used. The proofs, however, are rather more complicated.

It would be asserted that the estimate (2.40), which is referred to as the natural estimate, had the same asymptotic mean as $g_X^{(1)}(x)$, the estimate used with complete data, and in particular was asymptotically unbiased. It would also be stated that the asymptotic variance of the estimate was given by (2.21), with p and $g_Z(x)$ replaced by their average values, $A + (1 - A)p$ and $Ag_X(x) + (1 - A)g_Z(x)$, respectively. Here $g_Z(x)$ is the spectral density function of the stationary process $\{Z(t)\}$ given by $Z(t) = X(t)Y(t)$ for all t. Thus it would be assumed that the asymptotic variance of $g_X^{(4)}(x)$ was

$$n^{-1} W_n \{Ag_X(x) + (1 - A)g_Z(x)\}^2 d(x) / \{A + (1 - A)p\}^4 . \tag{2.41}$$

It is easy to verify that the natural estimate does indeed have the same asymptotic mean as $g_X^{(1)}(x)$. However, its variance is not given by (2.41). The correct formula is now derived.

Put $b_r = \{A + (1 - A)pu_r\}^{-1}$, $b = \{A + (1 - A)p^2\}^{-1}$ (not to be confused with a previous definition of b_r). Then

$$g_X^{(4)}(x) = g_1 + g_2 + g_3 + g_4 ,$$

where

$$2\pi n\, g_1 = \sum_{s,t=1}^{A_n} X(s)X(t)b_{s-t}W_{s-t}(x) ,$$

$$2\pi n\, g_2 = \sum_{s=1}^{A_n} \sum_{t=A_n+1}^{n} X(s)Z(t)b_{s-t}W_{s-t}(x) ,$$

$$2\pi n\, g_3 = \sum_{s=A_n+1}^{n} \sum_{t=1}^{A_n} Z(s)X(t)b_{s-t}W_{s-t}(x) ,$$

$$2\pi n\, g_4 = \sum_{s,t=A_n+1}^{n} Z(s)Z(t)b_{s-t}W_{s-t}(x) .$$

It is easily shown that $\sum|b_r - b| < \infty$. An argument similar to that which led to (2.21) shows that

$$\mathrm{Var}(g_1) = Ab^2n^{-1}W_n g_X(x)^2 d(x) + o(n^{-1}W_n) ,$$

$$\mathrm{Var}(g_4) = (1 - A)b^2n^{-1}W_n g_Z(x)^2 d(x) + o(n^{-1}W_n) .$$

We now show that the remaining terms of $\mathrm{Var}\{g_X^{(4)}(x)\}$ are negligible when compared with these.

First

$$\mathrm{Var}(g_2) = (2\pi n)^{-2} \sum_{q,s=1}^{A_n} \sum_{r,t=A_n+1}^{n} \Big[b_{q-r} w_{q-r}(x)\, b_{s-t} \overline{w_{s-t}(x)} \cdot \mathrm{Cov}\{X(q)Z(r), X(s)Z(t)\} \Big].$$

Now

$$\mathrm{Cov}\{X(q)Z(r), X(s)Z(t)\} = p u_{r-t} \{ Q_X(r-q, s-q, t-q) + R_X(s-q)R_X(t-r) + R_X(t-q)R_X(s-r) \} + p(u_{r-t} - p) R_X(r-q) R_X(t-s),$$

and $\mathrm{Var}(g_2)$ similarly splits up into four sums, T_1, T_2, T_3 and T_4, say. As in Section 2.3, T_1 is easily shown to be $O(n^{-1})$, and

$$|T_3| \leq Bn^{-2} \sum_{q,s=1}^{A_n} \sum_{r,t=A_n+1}^{n} |R_X(t-q)R_X(s-r)|$$

$$\leq Bn^{-2} \sum_{j,k=-\infty}^{\infty} |jR_X(j)\ kR_X(k)|$$

$$= O(n^{-2})$$

for some $B < \infty$, since $\{b_r\}$ is bounded. Similarly, we may show that $T_4 = O(n^{-2})$. Also, for some $B^* < \infty$,

$$|T_2| \leq B^* n^{-2} \sum_{q,s=1}^{A_n} \sum_{r,t=A_n+1}^{n} |w_{q-r}(x) \overline{w_{s-t}(x)} R_X(s-q) R_X(t-r)|$$

and using the Cauchy-Schwartz inequality, we see that

$$|T_2| \leq B^* n^{-1}\left\{n^{-1}\sum_{q,s=1}^{A_n}\sum_{r,t=A_n+1}^{n}|w_{q-r}(x)|^2|R_X(s-q)R_X(t-r)|\right\}^{\frac{1}{2}}$$

$$\cdot\left\{n^{-1}\sum_{q,s=1}^{A_n}\sum_{r,t=A_n+1}^{n}|w_{s-t}(x)|^2|R_X(s-q)R_X(t-r)|\right\}^{\frac{1}{2}}.$$

But

$$n^{-1}\sum_{q,s=1}^{A_n}\sum_{r,t=A_n+1}^{n}|w_{q-r}(x)|^2|R_X(s-q)R_X(t-r)|$$

$$\leq \sum_{|j|<A_n}\sum_{|k|<n-A_n}|R_X(j)R_X(k)|\; n^{-1}\sum_{\ell=-\infty}^{\infty}\max(|\ell|,n)|w_\ell(x)|^2$$

$$= o(W_n)$$

by (1.4) and (1.14), and clearly the same holds for the second factor of the upper bound for $|T_2|$. Therefore $T_2 = o(n^{-1}W_n)$, and hence $\mathrm{Var}(g_2) = o(n^{-1}W_n)$. A similar argument shows that $\mathrm{Var}(g_3)$ and $\mathrm{Cov}(g_2,g_3)$ are also both $o(n^{-1}W_n)$.

Secondly

$$\mathrm{Cov}(g_1,g_2) = (2\pi n)^{-2}\sum_{q,r,s=1}^{A_n}\sum_{t=A_n+1}^{n}\Big[b_{q-r}w_{q-r}(x)b_{s-t}\overline{w_{s-t}(x)}$$

$$\cdot\mathrm{Cov}\{X(q)X(r),X(s)Z(t)\}\Big]$$

$$= (2\pi n)^{-2} \sum_{q,r,s=1}^{A_n} \sum_{t=A_n+1}^{n} \Big[b_{q-r} W_{q-r}(x) b_{s-t} \overline{W_{s-t}(x)}$$

$$\cdot p\{Q_X(r-q, s-q, t-q) + R_X(s-q)R_X(t-r) + R_X(t-q)R_X(s-r)\}\Big] .$$

This splits up into three sums, each of which is easily seen to be $O(n^{-1})$. Similarly $\mathrm{Cov}(g_1, g_3)$, $\mathrm{Cov}(g_2, g_4)$ and $\mathrm{Cov}(g_3, g_4)$ are all $O(n^{-1})$.

Lastly

$$\mathrm{Cov}(g_1, g_4) = (2\pi n)^{-2} \sum_{q,r=1}^{A_n} \sum_{s,t=A_n+1}^{n} \Big[b_{q-r} W_{q-r}(x) b_{s-t} \overline{W_{s-t}(x)}$$

$$\cdot pu_{s-r}\{Q_X(r-q, s-q, t-q) + R_X(s-q)R_X(t-r)$$

$$+ R_X(t-q)R_X(s-r)\}\Big] .$$

Here the first sum may be shown to be $O(n^{-1})$, and the others to be $O(n^{-2})$.

Thus we have proved that the correct formula for the asymptotic variance of $g_X^{(4)}(x)$ in this situation is

$$\mathrm{Var}\{g_X^{(4)}(x)\} = \frac{n^{-1}W_n\{A g_X(x)^2 + (1-A) g_Z(x)^2\} d(x)}{\{A + (1-A)p^2\}^2} + o(n^{-1}W_n), \tag{2.42}$$

rather than (2.41).

It follows that the ratio of the true asymptotic variance to the asserted value (2.41) is

$$r = \frac{\{A + (1 - A)p\}^4 \{A g_X(x)^2 + (1 - A) g_Z(x)^2\}}{\{A + (1 - A)p^2\}^2 \{A g_X(x) + (1 - A) g_Z(x)\}^2}$$

This can in general take values both greater than or less than one, so that no general statement can be made about the values of r. However, if $g_X(x)$ is a constant, it reduces to

$$r = \{A + (1 - A)p\}^2 / \{A + (1 - A)p^2\} < 1.$$

Hence in this case, the asymptotic variance would be asserted to be higher than it is. It is expected that this would also be true if $g_X(x)$ had only small fluctuations about a constant value. However, this refers to the behaviour of $g_X(x)$ over the whole interval $(-\pi, \pi)$. It would not suffice to use the argument that since $g_X(x)$ is continuous, it is locally approximately constant.

Thus, provided $g_X(x)$ is fairly constant over its whole range, the effect of analysing this data as if the deletion mechanism were stationary is to over-estimate the asymptotic variance of the estimate by a factor which is approximately $\{A + (1 - A)p^2\}/\{A + (1 - A)p\}^2$. This factor is close to one whenever A, $(1 - A)$ or $(1 - p)$ is small.

In constructing the estimate $g_X^{(4)}(x)$, no use is made of the fact that part of the record is complete and part incomplete. Indeed, this estimate first arose in the discussion of an analysis of the data under the assumption that there

was no complete part. It seems plausible that a better estimate could be constructed using this information, if it were available. For example, asymptotically unbiased estimates of $g_X(x)$ could be constructed from each part, such as

$$g_X^{(c)}(x) = (2\pi A_n)^{-1} \sum_{s,t=1}^{A_n} X(s)X(t)w_{s-t}(x),$$

$$g_X^{(i)}(x) = \{2\pi(n-A_n)\}^{-1} \sum_{s,t=A_n+1}^{n} Z(s)Z(t)(pu_{s-t})^{-1}w_{s-t}(x).$$

It may be shown that

$$\mathrm{Var}\{g_X^{(c)}(x)\} = (An)^{-1}W_n g_X(x)^2 d(x) + o(n^{-1}W_n),$$

$$\mathrm{Var}\{g_X^{(i)}(x)\} = \{(1-A)n\}^{-1}W_n p^{-4} g_Z(x)^2 d(x) + o(n^{-1}W_n),$$

$$\mathrm{Cov}\{g_X^{(c)}(x), g_X^{(i)}(x)\} = o(n^{-1}W_n),$$

by methods similar to those used in deriving the asymptotic variance of $g_X^{(4)}(x)$.

These separate estimates could then be combined linearly as $\theta g_X^{(c)}(x) + (1-\theta)g_X^{(i)}(x)$. The choice of θ to minimise the variance of this combination is

$$\theta = Ag_Z(x)^2/\{p^4(1-A)g_X(x)^2 + Ag_Z(x)^2\}.$$

This involves the ratio $g_Z(x)/g_X(x)$, which is unknown, and p, which may be known, or can at least be estimated easily.

Suppose that p is, in fact, known. Noting that

$$\int_{-\pi}^{\pi} g_Z(x)\,dx = p\int_{-\pi}^{\pi} g_X(x)\,dx,$$

the ratio $g_Z(x)/g_X(x)$ could be approximated by p. This is, of course, exact when $g_X(x)$, and hence $g_Z(x)$, is constant. The linear combination is then

$$\left\{Ag_X^{(c)}(x) + (1-A)p^2 g_X^{(i)}(x)\right\}/\left\{A + (1-A)p^2\right\},$$

the asymptotic variance of which is

$$n^{-1}W_n\left\{Ag_X(x)^2 + (1-A)g_Z(x)^2\right\}d(x)/\left\{A + (1-A)p^2\right\}^2.$$

This is identical with (2.42), the asymptotic variance of $g_X^{(4)}(x)$.

It follows that, at least with this procedure, no asymptotic improvement results from the use of the information that part of the data is complete. As has been mentioned earlier, this does not imply that nothing is gained in the small sample case. It does, however, prove that there is no choice of θ depending on p but not on $g_Z(x)/g_X(x)$ which gives an estimate with a uniformly lower asymptotic variance than that of $g_X^{(4)}(x)$. For when $g_X(x)$ is constant, the best choice of θ leads to the same asymptotic variance as that of $g_X^{(4)}(x)$.

A more general nonstationary amplitude modulating process would arise if several different stationary amplitude

modulating processes were applied to different sections of the record. The results in this case are simple generalisations of those given above. In particular, if $g_X(x)$ is constant, it is again true that the asserted asymptotic variance of the natural estimate is higher than the true value. It is also true that if asymptotically unbiased estimates of $g_X(x)$ are constructed from each section and combined linearly as above, then the asymptotic variance of the combination is the same as that of the natural estimate.

CHAPTER 3. PERIODICALLY MISSING OBSERVATIONS

3.1 Introduction

In the last chapter some results were found for the situation where observations are missed in a fairly general random way. The assumptions made about the amplitude modulating process excluded the possibility of any periodic effect, since that process was weakly stationary and with an absolutely continuous spectral distribution function.

Two cases in which a particular type of periodic effect occurs are discussed in this chapter. The particular effect is where a fixed pattern of observations is missed repeatedly. This is a completely deterministic mechanism. In the first case this mechanism alone is responsible for all missed observations, while in the second case a random amplitude modulating process of the type described in Chapter 2 is in action as well.

The amplitude modulating function of this type of periodic mechanism is a nonrandom sequence of period h, say, denoted by $\{y_t : t = 0, \pm 1, \ldots\}$. The only assumptions made about this sequence are that $y_t = 0$ or 1 for each t, so that $y_t^2 = y_t$, and that for $r = 1, \ldots, h$,

$$R_y(r) = h^{-1} \sum_{t=1}^{h} y_t y_{t+r} > 0. \qquad (3.1)$$

This notation should not be confused with

$R_X(r) = \mathrm{Cov}\{X(t), X(t+r)\}$, as $\{y_t\}$ is a deterministic sequence. However, $R_y(r)$ can be thought of as a covariance sequence, since for $r > 0$,

$$R_y(r) = \lim_{t\to\infty} t^{-1} \sum_{s=1}^{t-r} y_s y_{s+r} .$$

Note that (3.1) implies a positive lower bound for $hR_y(0)$, the number of observations made in each pattern of length h. For $hR_y(r)$ is integer-valued, and hence (3.1) implies that $hR_y(r) \geq 1$.

But then

$$\{hR_y(0)\}^2 = \left(\sum_{t=1}^{h} y_t\right)^2$$

$$= \sum_{r=0}^{h-1} hR_y(r)$$

$$= hR_y(0) + \sum_{r=1}^{h-1} hR_y(r)$$

$$\geq hR_y(0) + h - 1,$$

which implies that $hR_y(0) \geq \frac{1}{2}\{1 + (4h-3)^{\frac{1}{2}}\}$. This is in fact a best lower bound for $hR_y(0)$ when (3.1) holds for $h \leq 19$, although apparently not for all higher values of h.

Note also that

$$0 \leq h^{-1} \sum_{t=1}^{h} (1 - y_t)(1 - y_{t+r}) = 1 - 2R_y(0) + R_y(r),$$

and hence $R_y(r) \geq 2R_y(0) - 1$. Thus a sufficient condition for (3.1) to hold is $R_y(0) \geq \frac{1}{2}$.

3.2 Periodically Missing Observations

In this section, a natural estimate of the spectral density of the process under observation is constructed when only the periodic mechanism is in effect. The estimate is the same as that used in a particular case by Jones (1962). The asymptotic variance and covariances of this estimate are derived; the result differs slightly from that given by Jones and Parzen (1963).

In order to avoid too much conflict of notation, suppose for the moment that $\{Z(t)\}$ is the process under observation, where $\{Z(t)\}$ satisfies (1.4) and (1.5), that $\{V(t)\}$ is the amplitude modulated process, where $V(t) = y_t Z(t)$, and that it is desired to estimate $g_Z(x)$ from $\{V(t) : t = 1, \ldots, n\}$.

The first step in the construction of an estimate of $g_Z(x)$ is to find an estimate of $R_Z(r)$. If $c_V(r)$ is defined as in (1.7), that is

$$c_V(r) = n^{-1} \sum_{t=1}^{n-|r|} V(t)V(t+|r|),$$

then

$$E\{c_V(r)\} = R_Z(r)n^{-1} \sum_{t=1}^{n-|r|} y_t y_{t+|r|}$$

$$= R_Z(r)R_y(r) + O(n^{-1}|r|).$$

It has been shown that $R_y(r)$ is bounded away from zero. Hence

$$E\{R_y(r)^{-1}c_V(r)\} = R_Z(r) + O(n^{-1}|r|),$$

and so $R_y(r)^{-1}c_y(r)$ is a suitable estimate of $R_Z(r)$. Thus a natural estimate of $g_Z(x)$ is

$$g_Z^{(5)}(x) = (2\pi)^{-1} \sum_{|r|<n} R_y(r)^{-1}c_y(r)w_r(x)$$
$$= (2\pi n)^{-1} \sum_{s,t=1}^{n} R_y(s-t)^{-1}y_s y_t Z(s)Z(t)w_{s-t}(x). \tag{3.2}$$

Now $\{y_t\}$ is a periodic sequence, so that

$$y_t = \sum_{j=1}^{h} G_y(j)e^{(i\phi_j t)}, \tag{3.3}$$

where $\phi_j = 2\pi j/h$, and

$$G_y(j) = h^{-1} \sum_{t=1}^{h} y_t e^{-(i\phi_j t)}. \tag{3.4}$$

This notation should not be confused with $G_X(x)$, the spectral distribution function. Note that the argument j is an integer. It is clear that $\{G_y(j)\}$ is a periodic sequence, $G_y(0) = R_y(0)$ and $G_y(-j) = \overline{G_y(j)}$.

Also $\{R_y(r)^{-1}\}$ is a periodic sequence, and so

$$R_y(r)^{-1} = \sum_{j=1}^{h} W_y(j)e^{(i\phi_j r)}, \tag{3.5}$$

where

$$W_y(j) = \sum_{r=1}^{h} R_y(r)^{-1}e^{-(i\phi_j r)}. \tag{3.6}$$

Hence $\{W_y(j)\}$ is a real, symmetric sequence, since $R_y(-r) = R_y(r)$.

If the expressions (3.3) and (3.5) for y_s, y_t and $R_y(s-t)^{-1}$ are substituted in (3.2), it becomes

$$g_Z^{(5)}(x) = (2\pi n)^{-1} \sum_{s,t=1}^{n} \sum_{j,k,\ell=1}^{h} \Big[W_y(j) e^{\{i(s-t)\phi_j\}} \cdot G_y(k) e^{(is\phi_k)} \overline{G_y(\ell)} e^{-(it\phi_\ell)} Z(s) Z(t) w_{s-t}(x) \Big],$$

which simplifies to

$$g_Z^{(5)}(x) = \sum_{j,k=1}^{h} H_y(j,k) g_{j,k}(x), \tag{3.7}$$

where

$$H_y(j,k) = \sum_{\ell=1}^{h} W_y(\ell) G_y(\ell-j) \overline{G_y(\ell-k)}, \tag{3.8}$$

$$g_{j,k}(x) = (2\pi n)^{-1} \sum_{s,t=1}^{n} Z(s) Z(t) w_{s-t}(x) e^{i(s\phi_j - t\phi_k)}. \tag{3.9}$$

It may be verified from (3.4), (3.6) and (3.8) that

$$H_y(j,j) = \begin{cases} 1 & \text{if } j \equiv 0 \pmod{h}, \\ 0 & \text{otherwise,} \end{cases}$$

that $\overline{H_y(j,k)} = H_y(k,j) = H_y(-j,-k)$ and that $H_y(j,k)$ is periodic in j and k.

In order to find the variance and covariances of $g_Z^{(5)}(x)$, consider

$$\operatorname{Cov}\{g_{j,k}(x_1), g_{\ell,m}(x_2)\}$$

$$= (2\pi n)^{-2} \sum_{q,r,s,t=1}^{n} \Big[w_{q-r}(x_1) \overline{w_{s-t}(x_2)} \exp\{i(q\phi_j - r\phi_k - s\phi + t\phi_m)\} \cdot \{Q_Z(r-q, s-q, t-q) + R_Z(s-q) R_Z(t-r) + R_Z(t-q) R_Z(s-r)\} \Big]$$

$$= T_1 + T_2 + T_3,$$

say. It is easy to show that $T_1 = O(n^{-1})$. Also, putting $s - q = a$, $t - r = b$, $q - r = c$, $r = d$, we see that

$$T_2 = (2\pi n)^{-2} \sum_{a,b,c} \Big[R_Z(a) R_Z(b) w_c(x_1) \overline{w_{a-b+c}(x_2)} \cdot \exp\{i(-a\phi + b\phi_m - c\phi_{\ell-j})\} \sum_d \exp(id\phi_{j-k-\ell+m}) \Big],$$

where the last sum is over the range

$$1 - \min(0, a+c, b, c) \leq d \leq n - \max(0, a+c, b, c),$$

which is taken to be void if the upper limit is less than the lower limit. If $j - k - \ell + m \not\equiv 0 \pmod h$, then the sum over d is uniformly bounded, whence $T_2 = O(n^{-1})$. If $j - k - \ell + m \equiv 0 \pmod h$, then

$$T_2 = (2\pi n)^{-2} \sum_{a,b,c} \Big[R_Z(a) R_Z(b) w_c(x_1) \overline{w_{a-b+c}(x_2)} \cdot \exp\{i(-a\phi + b\phi_m - c\phi_{\ell-j})\} N_{a,b,c} \Big],$$

where

$$N^*_{a,b,c} = n - \max(|a|, |b|, |c|, |a-c|, |b-c|, |a-b+c|),$$

$$N_{a,b,c} = \max(N^*_{a,b,c}, 0).$$

Note that

$$0 \leq n - N_{a,b,c} \leq |a| + |b| + |c|,$$

$$0 \leq n - N_{a,b,c} \leq |a-b+c| + |a| + |b|.$$

Thus, using (1.14) and the Cauchy-Schwarz inequality, it may be shown that

$$T_2 = (2\pi)^{-2} n^{-1} \sum_{a,b,c} \Big[R_Z(a) R_Z(b) w_c(x_1) \overline{w_{a-b+c}(x_2)}$$

$$\cdot \exp\{i(-a\phi_\ell + b\phi_m - c\phi_{\ell-j})\}\Big] + o(n^{-1}W_n)$$

$$= 2\pi n^{-1} \int_{-\pi}^{\pi} \{g_Z(y+\phi_\ell) g_Z(y+\phi_m) w(y+\phi_{\ell-j}; x_1) \overline{w(y;x_2)}\} dy$$

$$+ o(n^{-1}W_n),$$

where $w(y;x)$ is extended periodically outside the interval $(-\pi,\pi)$.

Now (1.11) and (1.12) may be used to show that T_2 is $o(n^{-1}W_n)$ unless $x_1 - x_2 = \phi_{\ell-j}$, when

$$T_2 = 2\pi n^{-1} g_Z(x_1+\phi_j) g_Z(x_2+\phi_k) \int_{-\pi}^{\pi} w(y+\phi_{\ell-j}; x_1) \overline{w(y; x_1-\phi_{\ell-j})} dy$$

$$+ o(n^{-1}W_n)$$

$$= n^{-1}W_n g_Z(x_1+\phi_j) g_Z(x_1+\phi_k) + o(n^{-1}W_n)$$

by (1.11).

Similarly we may show that T_3 is $O(n^{-1})$ unless $j - k - \ell + m \equiv 0 \pmod h$. When this holds, T_3 is $o(n^{-1}W_n)$, unless $x_1 + x_2 = -\phi_{j+m}$, when

$$T_3 = n^{-1}W_n g_Z(x_1+\phi_j) g_Z(x_1+\phi_k) + o(n^{-1}W_n).$$

This completes the description of $\mathrm{Cov}\{g_{j,k}(x_1), g_{\ell,m}(x_2)\}$. If a function $d_0(x)$ is defined by $d_0(0) = 1$, $d_0(x) = 0$ otherwise, then

$$\mathrm{Cov}\{g_{j,k}(x_1), g_{\ell,m}(x_2)\} = \Big[n^{-1}W_n g_Z(x_1+\phi_j) g_Z(x_1+\phi_k)$$

$$\cdot d_0(j-k-\ell+m)\{d_0(x_1-x_2+\phi_{\ell-j}) + d_0(x_1+x_2+\phi_{j+m})\}\Big]$$

$$+ o(n^{-1}W_n).$$

From (3.7) it follows that

$$\mathrm{Cov}\{g_Z^{(5)}(x_1), g_Z^{(5)}(x_2)\} = n^{-1}W_n \sum_{j,k=1}^{h} \Big[g_Z(x_1+\phi_j) g_Z(x_1+\phi_k) \cdot \sum_{m=1}^{h} H_y(j,k)\overline{H_y(j-k+m,m)}\{d_o(x_1-x_2+\phi_{k-m}) + d_o(x_1+x_2+\phi_{j+m})\}\Big] + o(n^{-1}W_n).$$

This may be stated as follows :

$$\mathrm{Cov}\{g_Z^{(5)}(x_1), g_Z^{(5)}(x_2)\} = n^{-1}W_n \sum_{j,k=1}^{h} \Big[g_Z(x_1+\phi_j) g_Z(x_1+\phi_k) H_y(j,k) \cdot \{\overline{H_y(j+q,k+q)} + \overline{H_y(j+r,k+r)}\}\Big] + o(n^{-1}W_n), \tag{3.10}$$

where the term $\overline{H_y(j+q,k+q)}$ arises if $x_1 - x_2 = \phi_q$, and the term $\overline{H_y(j+r,k+r)}$ arises if $x_1 + x_2 = \phi_r$.

In particular,

$$\mathrm{Var}\{g_Z^{(5)}(x)\} = n^{-1}W_n \sum_{j,k=1}^{h} \Big[g_Z(x+\phi_j) g_Z(x+\phi_k) \{|H_y(j,k)|^2 + H_y(j,k)\overline{H_y(j+r,k+r)}\}\Big] + o(n^{-1}W_n), \tag{3.11}$$

where the term $\overline{H_y(j+r,k+r)}$ is included only if $2x = \phi_r$. Thus when $2x \neq \phi_r$ for any r,

$$\mathrm{Var}\{g_Z^{(5)}(x)\} = n^{-1}W_n \sum_{j,k=1}^{h} g_Z(x+\phi_j) g_Z(x+\phi_k) |H_y(j,k)|^2 + o(n^{-1}W_n)$$

$$= n^{-1}W_n\Big\{g_Z(x)^2 + \sum_{j\neq k} g_Z(x+\phi_j)g_Z(x+\phi_k)\left|H_y(j,k)\right|^2\Big\} + o(n^{-1}W_n) \tag{3.12}$$

and if $x = 0$ or $\pm\pi$, then the asymptotic variance of $g_Z^{(5)}(x)$ is twice this value. However, if $2x = \phi_r$ for some $r \neq 0$ or $\pm h$, then (3.11) must be used, and may then be written

$$\mathrm{Var}\left\{g_Z^{(5)}(x)\right\} = n^{-1}W_n\Big(g_Z(x)^2 + \sum_{j\neq k}\Big[g_Z(x+\phi_j)g_Z(x+\phi_k)$$
$$\cdot\left\{\left|H_y(j,k)\right|^2 + H_y(j,k)\overline{H_y(j+r,k+r)}\right\}\Big]\Big) + o(n^{-1}W_n) \tag{3.13}$$

The exceptional behaviour at these values of x passes unnoticed by Parzen (1963), who states that (3.12) holds for all $x \neq 0, \pm\pi$.

Had the data been complete, the usual estimate $g_Z^{(1)}(x)$ would have been used, with

$$\mathrm{Var}\left\{g_Z^{(1)}(x)\right\} = n^{-1}W_n g_Z(x)^2 + o(n^{-1}W_n).$$

Now it may be shown that the sums in (3.12) and (3.13) are always positive, and hence the asymptotic variance is always increased by the absence of data.

Jones (1962) points out that

$$\sum_{j,k=1}^{h} g_Z(x+\phi_j)g_Z(x+\phi_k)\left|H_y(j,k)\right|^2$$
$$\leq \sup g_Z(x)^2 \sum_{j,k=1}^{h}\left|H_y(j,k)\right|^2,$$

and proposes the quantity

$$\bar{H}_y = \sum_{j,k=1}^{h} |H_y(j,k)|^2$$

as a measure of the increase in variance due to the absence of data. Parzen (1963) notes this may be simplified to

$$\bar{H}_y = h^{-2} \sum_{s,t=1}^{h} y_s^2 y_t^2 R_y(s-t)^{-2}. \qquad (3.14)$$

Since $y_t = 0$ or 1, this may be further simplified to

$$\begin{aligned} \bar{H}_y &= h^{-2} \sum_{s,t=1}^{h} y_s y_t R_y(s-t)^2 \\ &= h^{-2} \sum_{r,s=1}^{h} y_s y_{s+r} R_y(r)^{-2} \\ &= h^{-1} \sum_{r=1}^{h} R_y(r)^{-1} \\ &= W_y(0). \end{aligned} \qquad (3.15)$$

It has been shown that $R_y(r) \geq 2R_y(0) - 1$, and clearly $R_y(r) \leq R_y(0)$. Thus if $R_y(0) > \frac{1}{2}$, then

$$R_y(0)^{-1} \leq R_y(r)^{-1} \leq \{2R_y(0) - 1\}^{-1}, \qquad (3.16)$$

whence

$$\begin{aligned} R_y(0)^{-1} \leq \bar{H}_y &\leq (h-1)\{2R_y(0) - 1\}^{-1} + R_y(0)^{-1} \\ &\leq \{2R_y(0) - 1\}^{-1}. \end{aligned} \qquad (3.17)$$

It may be shown that the right hand inequality in (3.16) is exact if $y_t + y_{t+r} \geq 1$ for each t, and hence that the

first upper bound for $\bar{H}_y$ is exact when the pattern $\{y_1, \ldots, y_n\}$ contains only one zero, when $\bar{H}_y = (h^2 - h - 1)/\{(h-1)(h-2)\}$, a result given by Jones (1962).

One feature of the asymptotic variance of $g_X^{(5)}(x)$ is that it is discontinuous at each of the frequencies $\pm\emptyset_r$. The discontinuities at 0 and $\pm\pi$ are similar to those in the usual case, in that the value at each frequency is twice the value at adjacent frequencies. The other discontinuities are different, in that the value at a particular frequency may be less than at adjacent frequencies.

This may be seen as follows. Suppose that $g_Z(x) \equiv 1$. Then from (3.11),

$$\mathrm{Var}\{g_Z^{(5)}(x)\} = n^{-1}W_n \sum_{j,k=1}^{h} \{|H_y(j,k)|^2 + H_y(j,k)\overline{H_y(j+r,k+r)}\} + o(n^{-1}W_n), \tag{3.18}$$

where the term $H_y(j,k)\overline{H_y(j+r,k+r)}$ arises if $x = \pm\emptyset_r$. It has been shown that

$$\sum_{j,k=1}^{h} |H_y(j,k)|^2 = W_y(0),$$

and it may more generally be shown that

$$\sum_{j,k=1}^{h} H_y(j,k)\overline{H_y(j+r,k+r)} = W_y(r).$$

Thus the value of the asymptotic variance at $x = \pm\emptyset_r$ differs by $n^{-1}W_nW_y(r)$ from adjacent values.

Now it has also been stated that for $j \not\equiv 0 \pmod{h}$,

$$H_y(j,j) = \sum_{\ell=1}^{h} W_y(\ell) \left|G_y(\ell-j)\right|^2 = 0,$$

and hence at least one $W_y(r)$ must be negative, in general. Thus for at least one r, the asymptotic value of $\mathrm{Var}\left\{g_Z^{(5)}(\lambda\beta_r)\right\}$ is lower than adjacent values.

Many of the results of this section do not depend on the fact that $y_t = 0$ or 1. In particular, (3.10) and (3.11) hold whenever $\{y_t\}$ is a periodic sequence and (3.1) holds. Such general series cannot of course be interpreted as amplitude modulating functions. Later, however, the periodic mean value sequence of a random amplitude modulating process will be substituted in these formulae for the purposes of comparison with other results.

3.3 The Mixed Model

In the second model mentioned in Section 3.1, the deterministic periodic mechanism acts in combination with a stationary random mechanism of the type described in Chapter 2. It is assumed that the resulting amplitude modulating process is the product of the two separate ones. This is referred to as the mixed model. A more appealing way of combining two such amplitude modulating functions is described in Section 4.4, as a special case of the more general model of Chapter 4.

To be explicit, suppose that $\{X(t)\}$ is a process satisfying (1.4) and (1.5), that $\{y_t\}$ is a nonrandom periodic

sequence of zeros and ones of period h, satisfying (3.1), and that $\{Y(t)\}$ is a random amplitude modulating process satisfying (2.1) to (2.8). Let $Z(t) = X(t)Y(t)$ and $V(t) = y_t Z(t)$. Suppose also that it is desired to estimate $g_X(x)$ from $\{V(t) : t = 1, \ldots, n\}$.

First an estimate is constructed on the assumption that the sequence $\{pu_r R_y(r)\}$ is known, where p is defined by (2.2), u_r by (2.3) and $R_y(r)$ by (3.1). A natural estimate of $g_X(x)$ is then

$$g_X^{(6)}(x) = (2\pi)^{-1} \sum_{|r|<n} \{pu_r R_y(r)\}^{-1} c_V(r) w_r(x), \qquad (3.19)$$

where $c_V(r)$ is defined as in (1.7). This estimate has the usual property that its asymptotic mean is the same as that of $g_X^{(1)}(x)$.

To find the variance and covariances of $g_X^{(6)}(x)$, note that

$$\sum_r \left| \{pu_r R_y(r)\}^{-1} - p^{-2} R_y(r)^{-1} \right| < \infty, \qquad (3.20)$$

from (2.2), (2.5) and (3.1). But (3.19) may be written

$$\begin{aligned} g_X^{(6)}(x) &= p^{-2}(2\pi)^{-1} \sum_{|r|<n} R_y(r)^{-1} c_V(r) w_r(x) \\ &\quad + (2\pi)^{-1} \sum_{|r|<n} \left[\{pu_r R_y(r)\}^{-1} - p^{-2} R_y(r)^{-1} \right] c_V(r) w_r(x) \\ &= p^{-2} g_Z^{(5)}(x) + T, \end{aligned}$$

say, where $g_Z^{(5)}(x)$ is defined in (3.2). Now it was shown in

Section 2.2 that under the present assumptions, $\{Z(t)\}$ satisfies (1.3) and (1.4). Thus the conditions of the previous section are satisfied. It follows that the variance and covariances of $g_Z^{(5)}(x)$ are given by (3.10) and (3.11). Also, using (3.20) and an argument similar to that in Section 2.3, it may be shown that

$$\text{Var}\,(T) = O(n^{-1}),$$
$$\text{Cov}\{T, g_Z^{(5)}(x)\} = O(n^{-1}),$$

whence

$$\text{Var}\{g_X^{(6)}(x)\} = p^{-4}\,\text{Var}\{g_Z^{(5)}(x)\} + O(n^{-1}).$$

The argument uses no property of $w(y;x)$ other than the boundedness of $w_r(x)$. It may thus be easily extended to give

$$\text{Cov}\{g_X^{(6)}(x_1), g_X^{(6)}(x_2)\} = p^{-4}\,\text{Cov}\{g_Z^{(5)}(x_1), g_Z^{(5)}(x_2)\} + O(n^{-1}), \qquad (3.21)$$

where $\text{Cov}\{g_Z^{(5)}(x_1), g_Z^{(5)}(x_2)\}$ is given by (3.10).

The requirement that the sequence $\{pu_r R_y(r)\}$ be known may be relaxed, provided, of course, that (2.7), (2.8) and (3.1) hold. Suppose that $\underline{R}_y$ is a positive known lower bound for $R_y(r)$. Often $\{R_y(r)\}$ will be known, when $\underline{R}_y$ would be the lowest value in the periodic sequence $\{R_y(r)\}$. If $\{R_y(r)\}$ is unknown, then (3.1) implies that $R_y(r) \geq h^{-1}$. Hence some such $\underline{R}_y$ is always available. Since $c < pu_r$ for each r from (2.7), it follows that

$$pu_r R_y(r) > c\underline{R}_y .$$

An estimate of $pu_r R_y(r)$ is then $C_{yY}(r)$, where

$$C^*_{yY}(r) = (n - |r|)^{-1} \sum_{t=1}^{n-|r|} y_t Y(t) y_{t+r} Y(t+r), \qquad (3.22)$$

and

$$C_{yY}(r) = \begin{cases} c\underline{R}_y & \text{if } C^*_{yY}(r) \leq c\underline{R}_y , \\ C^*_{yY}(r) & \text{otherwise.} \end{cases} \qquad (3.23)$$

If, in (3.19), $pu_r R_y(r)$ is replaced by $C_{yY}(r)$, the estimate of $g_X(x)$ becomes

$$g^{(7)}_X(x) = (2\pi)^{-1} \sum_{|r|<n} C_{yY}(r)^{-1} c_V(r) w_r(x). \qquad (3.24)$$

It may now be shown that $g^{(7)}_X(x)$ has the same asymptotic mean, variance and covariances as $g^{(6)}_X(x)$. The result and the proof are very similar to those of Section 2.4, where it was shown that $g^{(2)}_X(x)$ and $g^{(3)}_X(x)$ had the same asymptotic mean, variance and covariances. Thus

$$\mathrm{Cov}\{g^{(7)}_X(x_1), g^{(7)}_X(x_2)\} = p^{-4} \mathrm{Cov}\{g^{(5)}_Z(x_1), g^{(5)}_Z(x_2)\} + o(n^{-1} W_n), \qquad (3.25$$

where $\mathrm{Cov}\{g^{(5)}_Z(x_1), g^{(5)}_Z(x_2)\}$ is given in (3.10).

The remarks made in Section 3.2 about the asymptotic variance of $g^{(5)}_Z(x)$ evidently also apply to that of $g^{(6)}_X(x)$. In particular, there are discontinuities at the frequencies $x = \pm\phi_r$. It is also true that the asymptotic

variance is always greater than it would have been had the data been complete.

CHAPTER 4. A MORE GENERAL PERIODIC MODEL

4.1 Assumptions

In this chapter, a fairly general model for a random amplitude modulating process with periodic effects is discussed. As one special case, it includes the mixed model of Section 3.3, which, although simple, is not very appealing. More trivially, it contains both the stationary random amplitude modulating process of Chapter 2 and the deterministic periodic mechanism of Chapter 3. Two other special cases are examined in Section 4.3 and 4.4.

The process under observation is again denoted by $\{X(t)\}$. The amplitude modulating process with the structure to be described will be denoted by $\{U(t)\}$, and the amplitude modulated process by $\{V(t)\}$ where $V(t) = U(t)X(t)$. This use of the symbol $\{V(t)\}$ is consistent with its use in Section 3.3, as the product amplitude modulating process $\{y_t Y(t)\}$ of that section possesses the structure to be described. The notation of Section 3.2 conflicts with this, but was adopted in order to make the results of Section 3.3 clearer.

The assumptions made about $\{U(t)\}$ are as follows:

$$U(t) \text{ is independent of } X(s); \tag{4.1}$$

$$P\{U(t) = 1\} = z_t = 1 - P\{U(t) = 0\}; \tag{4.2}$$

$$E\{U(t)U(t+r)\} = u_{t,r}; \tag{4.3}$$

$$E\{U(t)U(t+q)U(t+r)U(t+s)\} = v_{t,q,r,s}; \qquad (4.4)$$

z_t, $u_{t,r}$ and $v_{t,q,r,s}$ are each periodic in t with period h. (4.5)

These assumptions contrast with (2.1) to (2.4), in that the corresponding moments of $\{Y(t)\}$ were assumed to be independent of t.

The next assumption is that $R_z(r)$ is defined by and satisfies (3.1), that is

$$R_z(r) = h^{-1} \sum_{t=1}^{h} z_t z_{t+r} > 0.$$

Since z_t is not restricted to the values zero and one, the inequality derived from (3.1) no longer holds. A sufficient condition for (3.1) to hold is that $G_z(0) > \frac{1}{2}$, where $G_z(r)$ is defined as in (3.4).

Since $u_{t,r}$ is periodic in t, it follows that

$$u_{t,r} = \sum_{j=1}^{h} A_{j,r} e^{(it\phi_j)}, \qquad (4.6)$$

where

$$A_{j,r} = h^{-1} \sum_{t=1}^{h} u_{t,r} e^{-(it\phi_j)}. \qquad (4.7)$$

As in the case of the stationary model, the remaining assumptions depend on whether or not the values $A_{0,r}$ are known. In general, they would not be known.

If $A_{0,r}$ is known, it is assumed that

$$A_{0,r} > 0 \quad \text{for each } r, \qquad (4.8)$$

$$\sum_{q} |v_{t,r,q,q+s} - u_{t,r}\, u_{t+q,s}| \leq V(|r| + |s| + 1) < \infty \tag{4.9}$$

for each t. It may be shown that (4.8) and (4.9) imply that $A_{0,r}$ is bounded away from zero.

If $A_{0,r}$ is not known, it is assumed that

$$A_{0,t} > c \text{ for some known } c \text{ satisfying } 0 < c < R_z(t), \tag{4.10}$$

$$\sum_{q} |v_{t,r,q,s+q} - u_{t,r}\, u_{t+q,s}| \leq V < \infty \tag{4.11}$$

for each t.

Clearly (4.9) is satisfied whenever (4.11) is satisfied. If r and s are replaced in (4.9) by zero, it becomes

$$\sum_{q} |u_{t,q} - z_t\, z_{t+q}| < \infty . \tag{4.12}$$

It is easily verified that the product amplitude modulating process $\{y_t Y(t)\}$ of Section 3.3 satisfies these assumptions. If the moments of $\{U(t)\}$ do not depend on t, the structure of the stationary amplitude modulating process $\{Y(t)\}$ of Chapter 2 is recovered. If each z_t is either zero or one, then the purely periodic case of Chapter 3 arises. Thus the results about each of those models are contained in the results of this chapter.

4.2 The General Result

As before, assume first that the values $A_{0,r}$ are known. For

$$E\{c_V(r)\} = A_{0,r} R_X(r) + O(n^{-1}|r|),$$

and so

$$g_X^{(8)}(x) = (2\pi)^{-1} \sum_{|r|<n} A_{0,r}^{-1} c_V(r) w_r(x) \qquad (4.13)$$

is a natural estimate of $g_X(x)$, and has the same asymptotic mean as $g_X^{(1)}(x)$. It may be shown that

$$\sum_r |A_{0,r}^{-1} - R_Z(r)^{-1}| < \infty . \qquad (4.14)$$

Equation (4.13) may then be written

$$g_X^{(8)}(x) = (2\pi)^{-1} \sum_{|r|<n} R_Z(r)^{-1} c_V(r) w_r(x)$$

$$+ (2\pi)^{-1} \sum_{|r|<n} \{A_{0,r}^{-1} - R_Z(r)^{-1}\} c_V(r) w_r(x)$$

$$= g(x) + T, \qquad (4.15)$$

say. As in Section 2.3, it may be shown that $\mathrm{Var}(T)$ and $\mathrm{Cov}\{T, g(x)\}$ are both $O(n^{-1})$. Hence

$$\mathrm{Var}\{g_X^{(8)}(x)\} = \mathrm{Var}\{g(x)\} + O(n^{-1}).$$

The same argument may be extended immediately to give

$$\mathrm{Cov}\{g_X^{(8)}(x_1), g_X^{(8)}(x_2)\} = \mathrm{Cov}\{g(x_1), g(x_2)\} + O(n^{-1}).$$

From (3.5), it follows that

$$g(x) = \sum_{j=1}^{h} W_Z(j) T_j(x),$$

where

$$T_j(x) = (2\pi)^{-1} \sum_{|r|<n} c_V(r) e^{(ir\phi_j)} w_r(x)$$

$$= (2\pi n)^{-1} \sum_{s,t=1}^{n} V(s)V(t) e^{\{i(s-t)\phi_j\}} w_{s-t}(x).$$

Now

$$\mathrm{Cov}\{T_j(x_1), T_k(x_2)\}$$

$$= (2\pi n)^{-2} \sum_{q,r,s,t=1}^{n} \Big[\exp\{i(q-r)\phi_j - i(s-t)\phi_k\} w_{q-r}(x_1) \overline{w_{s-t}(x_2)}$$

$$\cdot \{ Q_X(r-q, s-q, t-q)\, v_{q,r-q,s-q,t-q}$$

$$+ R_X(r-q) R_X(t-s) (v_{q,r-q,s-q,t-q} - u_{q,r-q}\, u_{s,t-s})$$

$$+ R_X(s-q) R_X(t-r) (v_{q,r-q,s-q,t-q} - u_{q,s-q}\, u_{r,t-r})$$

$$+ R_X(t-q) R_X(s-r) (v_{q,r-q,s-q,t-q} - u_{q,t-q}\, u_{r,s-r})$$

$$+ u_{q,s-q}\, R_X(t-q)\, u_{r,t-r}\, R_X(t-r)$$

$$+ u_{q,t-q}\, R_X(t-q)\, u_{r,s-r}\, R_X(s-r) \} \Big] . \qquad (4.16)$$

The first four terms of this are easily seen to be $O(n^{-1})$, using (1.4) and (4.9). From (4.6), the fifth may be written

$$T_5 = (2\pi n)^{-2} \sum_{\ell,m=1}^{h} \sum_{q,r,s,t=1}^{n} \Big[\exp\{i(q-r)\phi_j - i(s-t)\phi_k\}$$

$$\cdot w_{q-r}(x_1) \overline{w_{s-t}(x_2)}\, A_{\ell,s-q}\, R_X(s-q)$$

$$\cdot \overline{A_{m,t-r}}\, R_X(t-r)\, \exp\{i(q\phi_\ell - r\phi_m)\} \Big] .$$

By an argument similar to that used in Section 3.2, the terms of T_5 with $\ell \not\equiv m \pmod h$ have a sum which is $O(n^{-1})$.

Hence

$$T_5 = \sum_{m=1}^{h} T_{5,m} + o(n^{-1}),$$

where

$$T_{5,m} = (2\pi n)^{-2} \sum_{q,r,s,t=1}^{n} \Big[\exp\{i(q-r)\phi_j - i(s-t)\phi_k\}$$

$$\cdot w_{q-r}(x_1)\overline{w_{s-t}(x_2)}\, A_{m,s-q}\, R_X(s-q)$$

$$\cdot \overline{A_{m,t-r}}\, R_X(t-r) e^{\,i(q-r)\phi_m}\Big] .$$

Thus, as in Section 3.2,

$$T_{5,m} = 2\pi n^{-1} \int_{-\pi}^{\pi} |g_m(y+\phi_k)|^2\, w(y-\phi_{j-k+m};x_1)\overline{w(y;x_2)}\, dy$$

$$+ o(n^{-1}W_n),$$

where

$$g_m(y) = (2\pi)^{-1} \sum_r A_{m,r}\, R_X(r)\, e^{-iry} . \qquad (4.17)$$

From (1.11) and (1.12), it follows that $T_{5,m}$ is $o(n^{-1}W_n)$, unless $x_2 = x_1 + \phi_{j-k+m}$, when

$$T_{5,m} = n^{-1}W_n |g_m(x_1+\phi_{j+m})|^2 + o(n^{-1}W_n).$$

Thus

$$T_5 = n^{-1}W_n \sum_{m=1}^{h} |g_m(x_1+\phi_{j+m})|^2\, d_o(x_1-x_2+\phi_{j-k+m}) + o(n^{-1}W_n).$$

In a similar way, it may be shown that

$$T_6 = n^{-1}W_n \sum_{m=1}^{h} |g_m(x_1+\phi_{j+m})|^2\, d_o(x_1+x_2+\phi_{j+k+m}) + o(n^{-1}W_n).$$

It follows that

$$\mathrm{Cov}\{g_X^{(8)}(x_1), g_X^{(8)}(x_2)\} = n^{-1}W_n \sum_{j,k,m=1}^{h} \Big[W_z(j)W_z(k) \cdot |g_m(x_1+\phi_{j+m})|^2 \{d_o(x_1-x_2+\phi_{j-k+m}) + d_o(x_1+x_2+\phi_{j+k+m})\}\Big] + o(n^{-1}W_n)$$

$$= n^{-1}W_n \sum_{j,k=1}^{h} W_z(j) |g_j(x+\phi_{j+k})|^2 \{W_z(j+k+q) + W_z(j+k+r)\} + o(n^{-1}W_n), \qquad (4.18)$$

where the term $W_z(j+k+q)$ arises if $x_1 - x_2 = \phi_q$, and the term $W_z(j+k+r)$ arises if $x_1 + x_2 = \phi_r$.

In particular,

$$\mathrm{Var}\{g_X^{(8)}(x)\} = n^{-1}W_n \sum_{j,k=1}^{h} \Big[W_z(j) |g_k(x_1+\phi_{j+k})|^2 \cdot \{W_z(j+k) + W_z(j+k+r)\}\Big] + o(n^{-1}W_n), \qquad (4.19)$$

where the term $W_z(j+k+r)$ is included if $x = \frac{1}{2}\phi_r$.

As in the purely periodic case of Section 3.2, there are discontinuities in the asymptotic variance at the frequencies $x = \frac{1}{2}\phi_r$. The effect of the discontinuity is to double the asymptotic variance at $x = 0, \pm\pi$, but is more complicated at the other frequencies.

The expression (4.19) is in fact more similar to (3.11) than at first appears. For

$$g_k(y) = (2\pi)^{-1} \sum_r A_{k,r} R_X(r) e^{-iry}$$

$$= (2\pi)^{-1} \sum_r h^{-1} \sum_{t=1}^{h} u_{t,r} R_X(r) e^{-iry} e^{-(it\phi_k)}$$

$$= (2\pi)^{-1} \sum_r h^{-1} \sum_{t=1}^{h} z_t z_{t+r} R_X(r) e^{-iry} e^{-(it\phi_k)}$$

$$+ (2\pi h)^{-1} \sum_r \sum_{t=1}^{h} (u_{t,r} - z_t z_{t+r}) R_X(r) e^{-iry} e^{-(it\phi_k)}$$

$$= \sum_{m=1}^{h} G_z(m) \overline{G_z(m-k)}\, g_X(y+\phi_{m-k}) + \sum_{t=1}^{h} h_t(y) e^{-(it\phi_k)} , \tag{4.20}$$

where

$$h_t(y) = (2\pi)^{-1} \sum_r (u_{t,r} - z_t z_{t+r}) R_X(r) e^{-iry} . \tag{4.21}$$

If the first sum of (4.20) alone were substituted for $g_k(y)$ in (4.18) or (4.19), the result would simplify to (3.10) or (3.11), respectively, which is the form that the corresponding covariances or variance would take if each U(t) were replaced by its mean z_t. Thus the effect of the second sum in (4.20) represents the effect of the variability of U(t).

It would be expected that this effect would be to increase the variance of the estimate. It may be verified that this is so in the mixed model of Section 3.3, and it is shown later that it is also true for one of the other examples of the present model. It has not, however, been proved in general.

In a now familiar way, the assumption that the values

$A_{0,r}$ are known may be relaxed. An estimate of $A_{0,r}$ is $C_u(r)$, defined precisely as in (2.22) and (2.23), using the quantity c defined in (4.10). This is substituted for $A_{0,r}$ in (4.13), to give an estimate of $g_X(x)$ with the same asymptotic mean and variance as $g_X^{(8)}(x)$.

4.3 Independent Omissions

One example of the model described in Section 4.1 is where $U(s)$ and $U(t)$ are independent for $s \neq t$. This is a periodic generalisation of the model examined by Scheinok (1965). In Scheinok's model, the probability of missing an observation was assumed to be constant.

In the periodic generalisation, for $r \neq 0$

$$u_{t,r} = z_t z_{t+r} \tag{4.22}$$

and so $A_{0,r} = R_z(r)$ for $r \neq 0$, while $A_{0,0} = G_z(0) \geqslant R_z(0)$. Thus (3.1) and a known lower bound for $R_z(r)$ imply (4.8) and (4.10). Note that since z_t can take any value between zero and one, (3.1) does not imply that such a bound is known. It is easy to verify that (4.11), and hence (4.9), is satisfied in this case, as the sum involves only a finite number of non-zero terms. Thus the assumptions of Section 4.1 are satisfied, and the results of Section 4.2 apply.

The function $h_t(y)$ defined in (4.17) now satisfies

$$h_t(y) = (2\pi h)^{-1} R_X(0) z_t(1 - z_t) \tag{4.23}$$

by (4.22), and is thus a constant. From (4.19), if $x \neq s\phi_r$ for any r, then

$$\mathrm{Var}\left\{g_X^{(s)}(x)\right\} = n^{-1} W_n \sum_{j,k=1}^{h} \Big[W_z(j) W_z(j+k) \cdot \Big| \sum_{m=1}^{h} \Big\{ G_z(m) \overline{G_z(m-k)}\, g_X(x+\phi_{j+m}) + (2\pi h)^{-1} R_X(0) \sum_{t=1}^{h} z_t(1-z_t) e^{-(it\phi_k)} \Big\} \Big|^2 \Big] + o(n^{-1} W_n). \tag{4.24}$$

When the squared modulus in (4.24) is expanded, the first term simplifies to (3.12), as has been noted earlier. The second term is

$$(2\pi nh)^{-1} W_n R_X(0) \sum_{j,k,m,t=1}^{h} \Big\{ W_z(j) W_z(j+k) G_z(m) \overline{G_z(m-k)} \cdot g_X(x+\phi_{j+m}) z_t(1-z_t) e^{(it\phi_k)} \Big\}$$

$$= (2\pi nh)^{-1} W_n R_X(0) \sum_m \Big[g_X(x+\phi_m) \sum_t \Big\{ z_t(1-z_t) \cdot \sum_j W_z(j) G_z(m-j) e^{-(it\phi_j)} \sum_k W_z(k) \overline{G_z(m-k)}\, e^{(it\phi_k)} \Big\} \Big]$$

$$= (2\pi nh)^{-1} W_n R_X(0) \sum_m \Big[g_X(x+\phi_m) \sum_t \Big\{ z_t(1-z_t) \cdot \Big| \sum_j W_z(j) G_z(m-j) e^{-(it\phi_j)} \Big|^2 \Big\} \Big]$$

$$\geqslant 0,$$

and the third term is similar. The fourth term is

$$(2\pi n)^{-2} W_n R_X(0)^2 \sum_{j,k,s,t=1}^{h} \left\{ W_z(j) W_z(j+k) \right.$$

$$\left. \cdot z_s(1 - z_s) z_t(1 - z_t) e^{-i(s-t)\phi_k} \right\}$$

$$= (2\pi h)^{-2} \sum_{s,t=1}^{h} \left\{ z_s(1 - z_s) z_t(1 - z_t) \right.$$

$$\left. \cdot \left| \sum_j W_z(j) e^{i(s-t)\phi_j} \right|^2 \right\}$$

$$\geqslant 0.$$

Thus the variance of the estimate is not less than it would have been if each $U(t)$ were replaced by its mean, z_t. The proof used here depends heavily on the assumption of independence of $U(s)$ and $U(t)$, and cannot be extended to the general model of Section 4.1.

4.4 Delayed Stationary Omissions

In the second example of the model of Section 4.1, a purely periodic mechanism and a general stationary random mechanism of the type described in Chapter 2 are combined. The way in which they are combined is more involved than the simple multiplication used in Section 3.3.

Suppose that a purely periodic mechanism results in the deletion of some of the observations. Give each remaining observation a subsidiary index, according to its position in the

sequence of remaining observations. Suppose now that extra observations are deleted by a random mechanism which is stationary with respect to the subsidiary index, rather than with respect to the original index, time. This could be described by saying that the random mechanism is delayed or suspended by the periodic mechanism.

The structure may be described formally as follows. Let $\{y_t\}$ be a periodic sequence of period h, with $y_t = 0$ or 1, such that (3.1) is satisfied. Suppose that $\{Y(t)\}$ is a random amplitude modulating process satisfying (2.1) to (2.8). Let

$$T_t = \sum_{s=1}^{t} y_s \tag{4.25}$$

and let $U(t) = y_t\, Y(T_t)$. It is clear that

$$E\{U(t)\} = y_t\, p = z_t,$$

say, is periodic, and

$$\begin{aligned} E\{U(t)U(t+r)\} &= y_t\, y_{t+r}\, E\{Y(T_t)Y(T_{t+r})\} \\ &= y_t\, y_{t+r}\, u(T_{t+r} - T_t) \\ &= u_{t,r}, \end{aligned}$$

where, in this equation, the symbol u_r defined in (2.3) is denoted by $u(r)$. Now if, for instance, $r > 0$, then

$$T_{t+r} - T_t = \sum_{s=t+1}^{t+r} y_s$$

is evidently periodic in t. The same is true for $r \leqslant 0$, and so $u_{t,r}$ is periodic in t. Thus the first two parts of (4.5) are satisfied. It is straightforward to verify that the last part of (4.5) is satisfied, and that the remaining conditions of Section 4.1 are implied by the conditions on $\{Y(t)\}$.

This type of deletion of data could arise as follows. Suppose that some apparatus is used to record data, and is liable to omit certain observations in a stationary random way. Suppose then that this apparatus is used to record n_1 data items, then the next n_2 items are allowed to pass unrecorded. If the failures of the apparatus are unaffected by the interruption of its use, then the overall amplitude modulating function would be of the type described.

The expression for the asymptotic variance of the estimate of $g_X(x)$ is not simple in this case. Thus this would probably not be a suitable model to fit to any data, unless there were definite reasons for believing that it did, in fact describe the situation exactly. Nevertheless, it does provide an example of how this general type of amplitude modulating process could arise in practice.

4.5 Applications

The assumptions made about the model of Section 4.1 involve many parameters. For certain moments are allowed to be

periodic functions of time, rather than constants. If all those moments are known then no penalty is incurred, apart from the complexity of the formula (4.17) for the asymptotic covariance of the spectral estimates.

In general, the moments will not be known. Then it will be necessary to use the modified estimate. It has been stated without proof that the modification does not affect the asymptotic mean or variance of the estimate. However, the rate of approach to the asymptotic formula may be worse, and also the estimation of quantities involved in the asymptotic formula may be difficult.

For these reasons it may be desirable to restrict the model in some way. The models of Chapters 2 and 3 are special cases, and could be used when suitable. The delayed stationary omissions model is given more as an example of the structure of Section 4.1 than as a usable model. The model of independent omissions, on the other hand, might well give an adequate representation of data for which the earlier models were inadequate. It has three attractive properties :

(i) all the quantities involved in the expression for the asymptotic variance may be found from the h values $\{z_1, \ldots, z_h\}$;

(ii) it can allow simultaneously for some of the observations to be made certainly and some to be missed certainly ;

(iii) the formula for the asymptotic variance simplifies as has been shown.

There are evidently other models which satisfy the assumptions of Section 4.1. One which has not been mentioned is the complement of the mixed model of Section 3.3 ; an amplitude modulating process $\{U(t)\}$ has this structure if the process $\{1 - U(t)\}$ has the structure of the mixed model. The discussion of this chapter is not intended to be exhaustive.

4.6 A Stationary Model

The random amplitude modulating process $\{Y(t)\}$ of Chapter 2 has been referred to for convenience as the stationary model. It is, however, not the only model which gives rise to a stationary amplitude modulating process. The models of Chapter 3 and Section 4.1 all give rise, in general, to nonstationary amplitude modulating processes. Stationary models may, however, be constructed from them.

Suppose, for example, that $\{U(t)\}$ satisfies the conditions of Section 4.1, and that T is uniformly distributed on $\{1, \ldots, h\}$, independently of all other random variables. Let $U_S(t) = U(t+T)$. Then it is easy to verify that $\{U_S(t)\}$ is weakly stationary, with mean $G_Z(0)$ and autocovariance sequence $\{A_{0,r} - G_Z(0)^2\}$. However, it may also be verified that $\{U_S(t)\}$ does not, in general, satisfy the conditions of Chapter 2.

The natural estimates of $g_X(x)$ from data which is incomplete due to the action of $\{U_S(t)\}$ are the same as if

it were $\{U(t)\}$. The same is also true of their asymptotic variances and covariances. It is clear that the expressions for these covariances involve more than the mean and spectrum of $\{U_S(t)\}$. This is in contrast with the results of Sections 2.3 and 2.4, in which only the mean and spectrum of $\{Y(t)\}$ were required. The difference is evidently due to the periodic effects present in $\{U_S(t)\}$.

Thus it is clear that once periodic effects are allowed, no general results can be obtained which involve only first and second order properties of the amplitude modulating process.

CHAPTER 5. OTHER ESTIMATES OF THE SPECTRAL DENSITY FUNCTION

5.1 Introduction

The natural estimate of the spectral density function, introduced in Section 1.3, is not the only estimate which may be used. In this chapter, two other ways of dealing with missing observations are described.

The first is the only example in the literature of an asymptotically unbiased estimate which is not a natural estimate. Unfortunately, its use seems to be restricted to a rather special situation. There does not appear to be any way of constructing such an estimate in more general cases.

The second gives rise to an asymptotically biased estimate, but is particularly simple to apply. It is shown that when only few observations are missed, the asymptotic bias is small. A modified estimate with a smaller asymptotic bias is produced, but it does not have the simplicity of the original estimate. Thus this procedure only appears useful when few observations are missing.

5.2 Neave's "Alias-improved" Estimate

The first example of an estimate of the spectral density function which is not a natural estimate is proposed in a recent paper by Neave (1970). The situation in which it may be used is where one has "initially scarce" data. That is,

the interval between observations has been reduced by some integral factor at some point during the time in which observations have been made. Thus, for instance, one might have a series which was, in the past, sampled yearly, but has recently been sampled quarterly.

Evidently, if sufficient data were recorded at the shorter interval, one could ignore the older data. However, if the change in observation interval occurred only relatively recently, it might be very inefficient to ignore the information contained in the older data.

One way to analyse such data would be to use a natural estimate based on all of the data. However, in this case Neave found that this estimate had certain disadvantages. The main ones are that it is difficult to make it nonnegative, and that there tend to be disturbances at certain frequencies. To be specific, if the final interval were one unit, and the original interval were h units, then there would tend to be disturbances at the frequencies $\frac{1}{2}\emptyset_r = \pi r/h$.

Another way of treating these data is as follows. The first part of the data gives information only about the covariances whose lags are multiples of h. This means that this part of the data gives information only about the h-aliased spectral density function

$$g_X^{[h]}(x) = \sum_{r=1}^{h} g_X(x+\emptyset_r) \qquad (5.1)$$

and not about the separate terms of the sum, for example

their ratios.

An estimate of $g_X(x)$ is constructed as follows. First an estimate, $g_1(x)$, of $g_X(x)$ is constructed in the usual way from the second, complete, part of the data. Then an estimate, $g_2(x)$, of $g_X^{[h]}(x)$ is constructed from the entire data. A second estimate, $g_3(x)$, of $g_X(x)$ is constructed to satisfy

a) its h-alias

$$\sum_{r=1}^{h} g_3(x+\phi_r) \tag{5.2}$$

equals $g_2(x)$, and

b) the ratios of the terms in the sum (5.2) equal, for each x, the corresponding ratios for $g_1(x)$.

Neave calls this the "alias-improved" estimate. It is easy to verify that it must be given by

$$g_3(x) = g_2(x)g_1(x)\left\{\sum_{r=1}^{h} g_1(x+\phi_r)\right\}^{-1}.$$

Because the estimates $g_1(x)$ and $g_2(x)$ may be constructed to be nonnegative, $g_3(x)$ may also be constructed to have this useful property. It is not so obvious that the alias-improved estimate does not share the other disadvantage of the natural estimate, the disturbance at the frequencies $\pm\phi_r$. However, Neave found in a simulation study that the alias-improved estimate was better in this respect. He also found that its variance was less than that of the natural estimate, although this was not indicated by an asymptotic result.

The situation in which this estimate is used is unlike most missing data situations. For part of the data are complete, and part are so incomplete that $g_X(x)$ cannot be estimated from that part alone. The "alias-improved" estimate might be generalised to some other situations with these properties, but it is difficult to see how it could be used in general.

5.3 Replacement of Missing Data

When there are only a few observations missing, it may be possible to replace them without seriously affecting the data. One would wish to use a value which was in some sense close to the missing value. For example, the value of some highly correlated series might be used, or a weighted average of neighbouring values in the original series. The latter method was used by Brillinger (1969) to complete some physical time series.

Since it is assumed that only few observations are missing, it seems reasonable to assume a random deletion mechanism. Suppose that $\{X_1(t)\}$ is the series under incomplete observation, and that $\{X_2(t)\}$ is the series used to fill the gaps, where both $\{X_1(t)\}$ and $\{X_2(t)\}$ satisfy (1.4) and (1.5), and are stationarily cross-correlated. That is

$$E\{X_1(t)X_2(t+r)\} = R_{1,2}(r) = R_{2,1}(-r),$$

independently of t. The subscripts 1 and 2, rather than

X_1 and X_2, are used to denote quantities associated with $\{X_1(t)\}$ and $\{X_2(t)\}$, respectively. Thus for instance, $\mathrm{Cov}\{X_1(t), X_1(t+r)\}$ is denoted by $R_1(r)$.

Suppose also that $\{Y(t)\}$ is a random amplitude modulating process satisfying (2.1) to (2.8). Let

$$\begin{aligned} Z(t) &= Y(t)X_1(t) + \{1 - Y(t)\}X_2(t) \\ &= Y(t)\{X_1(t) - X_2(t)\} + X_2(t) \end{aligned} \tag{5.3}$$

That is,

$$Z(t) = \begin{cases} X_1(t) & \text{when observed,} \\ X_2(t) & \text{otherwise.} \end{cases}$$

Under a suitable condition on the mixed fourth moments of $X_1(t)$ and $X_2(t)$, it is easy to check that $\{Z(t)\}$ satisfies (1.3) and (1.4), with

$$\begin{aligned} R_Z(r) &= pu_r\{R_1(r) + R_2(r) - R_{1,2}(r) - R_{2,1}(r)\} \\ &\quad + p\{R_{1,2}(r) + R_{2,1}(r)\} + (1 - 2p)R_2(r) \\ &= p^2 R_1(r) + (1 - p)^2 R_2(r) + p(1 - p)\{R_{1,2}(r) + R_{2,1}(r)\} \\ &\quad + p(u_r - p)\{R_1(r) + R_2(r) - R_{1,2}(r) - R_{2,1}(r)\}. \end{aligned}$$

Thus

$$\begin{aligned} g_Z(x) &= p^2 g_1(x) + (1 - p)^2 g_2(x) + 2p(1 - p)c(x) \\ &\quad + \int_{-\pi}^{\pi} \{g_1(y) + g_2(y) - 2c(y)\} a(x-y)\, dy , \end{aligned} \tag{5.4}$$

where

$$c(x) = (2\pi)^{-1} \sum_r \tfrac{1}{2}\{R_{1,2}(r) + R_{2,1}(r)\} e^{irx}$$

is the cospectral density function of $X_1(t)$ and $X_2(t)$, and

$$a(x) = (2\pi)^{-1} \sum_r p(u_r - p) e^{irx}.$$

Note that this sum converges by (2.9).

The spectral density function of $\{Z(t)\}$, $g_Z(x)$, may be estimated by $g_Z^{(1)}(x)$, defined as in (1.9), and $g_Z^{(1)}(x)$ is asymptotically unbiased, and has an asymptotic variance given by (1.15). However, $g_Z(x)$ and $g_1(x)$ are not, in general, identical, so that $g_Z^{(1)}(x)$ is an asymptotically biased estimate of $g_1(x)$. The asymptotic bias may be shown to be $O(1 - p)$ as $p \to 1$, so that if $1 - p$ were small, it may well be small enough to be ignored.

In the case where a weighted average is used to replace the missing value, the expression for $g_Z(x)$ simplifies somewhat. It is assumed that the average is chosen in such a way that all the values required are certain to be available. This is convenient, and can be defended on the grounds that when few observations are missing, they are likely to be fairly well separated. Thus

$$X_2(t) = \sum \alpha_r X_1(t-r)$$

where $\alpha_0 = 0$, and the sum is over a finite set. This leads to

$$c(x) = g_1(x) \operatorname{Re} \alpha(x),$$

$$g_2(x) = g_1(x)\,|\alpha(x)|^2,$$

where

$$\alpha(x) = \sum_r \alpha_r\, e^{irx},$$

and so

$$g_Z(x) = |p + (1-p)\alpha(x)|^2\, g_1(x) + \int_{-\pi}^{\pi} |1 - \alpha(y)|^2\, g_1(y) a(x-y)\, dy. \qquad (5.5)$$

Here $\mathrm{Re}\,\alpha(x)$ denotes the real part of the complex function $\alpha(x)$.

If $g_Z(x)$, $\alpha(x)$, $a(x)$ and p were known, (5.5) would be an integral equation of the Fredholm type for $g_1(x)$. Such equations cannot in general be solved in closed form, so that the elimination of the bias of $g_Z^{(1)}(x)$ as an estimate of $g_1(x)$ seems to be difficult.

However, as has been stated,

$$g_Z(x) = g_1(x) + O(q)$$

as $q = 1 - p \to 0$, which suggests that

$$g_1^*(x) = g_Z(x) + qh(x) \qquad (5.6)$$

may be an approximate solution of (5.5) for small q. Substituting in (5.5), we obtain

$$g_Z(x) = |(1-q) + q\alpha(x)|^2 \{g_Z(x) + qh(x)\} + q\int_{-\pi}^{\pi} |1 - \alpha(y)|^2 \{g_Z(y) + qh(y)\} \{a(x-y)/q\} dy,$$

where we write $a(x)$ as $q\{a(x)/q\}$ since

$$\int_{-\pi}^{\pi}\{a(x)/q\}dx = 1 - q \to 1$$

as $q \to 0$. Ignoring terms of order q^2,

$$g_Z(x) = \{1 - 2q + 2q \operatorname{Re} \alpha(x)\} g_Z(x) + qh(x) - q\int_{-\pi}^{\pi} |1 - \alpha(y)|^2 g_Z(y)\{a(x-y)/q\}dy$$

or

$$h(x) = 2\{1 - \operatorname{Re} \alpha(x)\} g_Z(x) - \int_{-\pi}^{\pi} |1 - \alpha(y)|^2 g_Z(y)\{a(x-y)/q\}dy.$$

The approximate solution becomes

$$g_1^*(x) = \{1 + 2q - 2q \operatorname{Re} \alpha(x)\} g_Z(x) - \int_{-\pi}^{\pi} |1 - \alpha(y)|^2 g_Z(y)a(x-y)\, dy$$

or

$$g_1^*(x) = |1 - q + q\alpha(x)|^{-2} g_Z(x) - \int_{-\pi}^{\pi} |1 - \alpha(y)|^2 g_Z(y)a(x-y)\, dy, \qquad (5.7)$$

where the latter expression for $g_1^*(x)$ differs from the former only in powers of q higher than the first. It may be verified that $\{g_1(x) - g_1^*(x)\}$ is $O(q^2)$ as $q \to 0$.

This suggests that if a new estimate of $g_1(x)$ is formed by substituting $g_Z^{(1)}(x)$ for $g_Z(x)$ in (5.7), the resulting estimate might have a lower asymptotic bias than $g_Z^{(1)}(x)$. It is certainly true for small enough q.

As in Section 2.3, the variance of this estimate is

$$\left|p + (1 - p)\alpha(x)\right|^{-4} \operatorname{Var}\left\{g_Z^{(1)}(x)\right\} + O(n^{-1}).$$

If the weights α_r correspond to a true average, that is if $\alpha_r \geqslant 0$ and $\alpha(0) = 1$, or more generally if $\sum |\alpha_r| \leqslant 1$, then

$$\left|p + (1 - p)\alpha(x)\right| < 1,$$

so that the variance of this estimate is greater than that of $g_Z^{(1)}(x)$. Thus the reduction in bias is achieved at the cost of an inflation of the variance.

Nothing has yet been said about the choice of $\alpha(x)$. Intuitively, it seems plausible that the average should be chosen to predict the missing value well, in some sense. It is known that the function $\alpha(x)$ corresponding to the best predictor in mean square is

$$\alpha^*(x) = 1 - 2\pi g_1(x)^{-1} / \int_{-\pi}^{\pi} g_1(y)^{-1}\, dy.$$

This does not in general give a finite average, and its use implies knowledge of $g_1(x)$. However, a finite average could be chosen on the basis of a preliminary estimate of $g_1(x)$. The optimal choice of $\alpha(x)$ has not been examined.

This section is not intended to be a full discussion of this method of analysing incomplete data. Certain comments may be made :

a) the computation involved in the modified estimate makes it less appealing than the original estimate ;

b) the bias involved in the original estimate would make it

unacceptable unless $1 - p$ were small ;

c) provided $1 - p$ was sufficiently small, the unmodified estimate would be very simple to use, as it is the classical estimate of the spectral density function of a slightly modified series.

CHAPTER 6. MULTIVARIATE DATA

6.1 Introduction

All the previous results have been for a univariate series $\{X(t)\}$. However, missing data problems arise also in the analysis of multivariate series, and deserve some attention. In this chapter the stationary amplitude modulating process of Chapter 2 is generalised to the multivariate case, and two special cases are examined. The periodic models of Chapters 3 and 4 could also be so extended, but they are already rather cumbersome. The general model of Chapter 4 is somewhat vague, and could only become vaguer by the extension. For these reasons, the discussion is confined to the stationary random model

Let $\{\underline{X}(t)\}$ be a weakly stationary h-variate process, $\underline{X}'(t) = \{X_1(t), \ldots, X_h(t)\}$.

Suppose

$$E\{X_j(t)\} = 0,$$

$$E\{X_j(t)X_k(t+r)\} = R_{j,k}(r) = R_{k,j}(-r).$$

Here and later the subscripts j,k,ℓ and m are assumed to be in the range $\{1, \ldots, h\}$.

Evidently $R_{j,j}(r)$ is the autocovariance sequence of the j-th component of $\{\underline{X}(t)\}$. The other values $R_{j,k}(r)$ are termed crosscovariances. As can be seen, they are not in

general symmetric. Let

$$g_{j,k}(x) = (2\pi)^{-1} \sum_r R_{j,k}(r)\ e^{-irx}.$$

If $j = k$, $g_{j,k}(x)$ is the spectral density function of the j^{th} component of $\{\underline{X}(t)\}$, and for $j \neq k$, $g_{j,k}(x)$ is a cross-spectral density function. Note that, in general, cross-spectral density functions are complex-valued, and that

$$\overline{g_{j,k}(x)} = g_{j,k}(-x) = g_{k,j}(x). \tag{6.1}$$

The real and imaginary parts of $g_{j,k}(x)$ are sometimes dealt with separately, as the cospectral and quadrature spectral density functions, respectively. However, it is more compact to treat them as a single complex function.

We make the following assumptions about $\{\underline{X}(t)\}$. Similarly to (1.4), the first is

$$\sum_r |rR_{j,k}(r)| < \infty. \tag{6.2}$$

Also, if

$$Q_{j,k,\ell,m}(q,r,s) = E\{X_j(t)X_k(t+q)X_\ell(t+r)X_m(t+s)\}$$
$$- R_{j,k}(q)R_{\ell,m}(s-t) - R_{j,\ell}(r)R_{m,k}(q-s)$$
$$- R_{j,m}(s)R_{k,\ell}(r-q),$$

then it is assumed that

$$\sum_{q,r,s} |Q_{j,k,\ell,m}(q,r,s)| < \infty, \quad j,k,\ell,m = 1, \ldots, h. \tag{6.3}$$

The basic result in the estimation of cross-spectral density functions is the following. Suppose that $w(y;x)$ is as in Section 1.2, and that

$$g_{j,k}^{(1)}(x) = (2\pi n)^{-1} \sum_{s,t=1}^{n} X_j(s) X_k(t) w_{t-s}(x). \tag{6.4}$$

Then for $x \neq 0, \pm\pi$,

$$\mathrm{Cov}\left\{g_{j,k}^{(1)}(x), g_{\ell,m}^{(1)}(x)\right\} = n^{-1} W_n g_{j,\ell}(x) g_{m,k}(x) + o(n^{-1} W_n) \tag{6.5}$$

while if $x = 0$ or $\pm\pi$, then

$$\mathrm{Cov}\left\{g_{j,k}^{(1)}(x), g_{\ell,m}^{(1)}(x)\right\}$$

$$= n^{-1} W_n \left\{g_{j,\ell}(x) g_{m,k}(x) + g_{j,m}(x) g_{\ell,k}(x)\right\} + o(n^{-1} W_n), \tag{6.6}$$

and if $x_1 \neq \pm x_2$, then

$$\mathrm{Cov}\left\{g_{j,k}^{(1)}(x_1), g_{\ell,m}^{(1)}(x_2)\right\} = o(n^{-1} W_n). \tag{6.7}$$

Estimates of the cospectral and quadrature spectral density functions can be constructed using (6.1). Their asymptotic variances and covariances can then be found from (6.5) to (6.7). This is the form in which these results are given in Rosenblatt (1959).

Note that for complex random variables

$$\mathrm{Cov}\left\{G_1, G_2\right\} = E\left[\left\{G_1 - E(G_1)\right\}\overline{\left\{G_2 - E(G_2)\right\}}\right].$$

6.2 A Multivariate Stationary Amplitude Modulating Process

Suppose that $\{\underline{Y}(t)\}$ is an h-variate process, satisfying

$$\left.\begin{aligned} &Y_j(t) \text{ is independent of } X_k(s); \\ &P\{Y_j(t) = 1\} = p_j = 1 - P\{Y_j(t) = 0\} \\ &\text{for some } p_j \text{ with } 0 < p_j \leq 1; \\ &E\{Y_j(t)Y_k(t+r)\} = u_{j,k}(r); \\ &E\{Y_j(t)Y_k(t+q)Y_\ell(t+r)Y_m(t+s)\} = v_{j,k,\ell,m}(q,r,s) \end{aligned}\right\} \quad (6.8)$$

for all q,r,s and t.

A process $\{\underline{Y}(t)\}$ satisfying these conditions is evidently a generalisation of the univariate process $\{Y(t)\}$ of Chapter 2. In particular, each component of $\{\underline{Y}(t)\}$ is of that form. The following assumptions are made.

If $u_{j,k}(r)$ is known for all r and for $j,k = 1,\ldots,h$, it is assumed that $u_{j,k}(r) > 0$, and

$$\sum_q |v_{j,k,\ell,m}(r,q,q+s) - u_{j,k}(r)u_{\ell,m}(s)| \leq V(|r| + |s| + 1) < \infty, \quad (6.9)$$

while if $u_{j,k}(r)$ is unknown, it is assumed that

$u_{j,k}(r) > c$ for some known c satisfying $0 < c < p_j\, p_k$, (6.10)

and that

$$\sum_q |v_{j,k,\ell,m}(r,q,q+s) - u_{j,k}(r)u_{\ell,m}(s)| < \infty. \quad (6.11)$$

It is clear from (6.9) or (6.11) with $r = s = 0$, $j = k$ and $\ell = m$ that

$$\sum_{q} |u_{j,\ell}(q) - p_j p_\ell| < \infty . \tag{6.12}$$

The amplitude modulated process is $\{\underline{Z}(t)\}$, where $Z_j(t) = X_j(t)Y_j(t)$, $j = 1, \ldots, h$. It is weakly stationary, and its cross-covariance function is

$$R^{(Z)}_{j,k}(r) = R_{j,k}(r)u_{j,k}(r).$$

The superscript (Z) is used to indicate that the quantity is derived from $\{\underline{Z}(t)\}$, rather than from $\{\underline{X}(t)\}$. The cross-spectral density functions of $\{\underline{Z}(t)\}$ are

$$g^{(Z)}_{j,k}(x) = p_j p_k\, g_{j,k}(x) + \int_{-\pi}^{\pi} g_{j,k}(y)a_{j,k}(x-y)\,dy, \tag{6.13}$$

where

$$a_{j,k}(x) = (2\pi)^{-1} \sum_{r} \{u_{j,k}(r) - p_j p_k\}\, e^{irx}.$$

It may be verified that $\{\underline{Z}(t)\}$ satisfies (6.1) and (6.2), so that $g^{(Z)}_{j,k}(x)$ may be estimated in the same way as $g_{j,k}(x)$ with asymptotic variances and covariances given by expressions like (6.5) to (6.7).

If $u_{j,k}(r)$ is known, a natural estimate of $g_{j,k}(x)$ is

$$g^{(2)}_{j,k}(x) = (2\pi n)^{-1} \sum_{s,t=1}^{n} Z_j(s)Z_k(t)u_{j,k}(t-s)^{-1}\, w_{t-s}(x). \tag{6.14}$$

As in previous work, (6.10) implies that

$$\mathrm{Cov}\left\{g^{(2)}_{j,k}(x_1), g^{(2)}_{\ell,m}(x_2)\right\}$$

$$= \mathrm{Cov}\left\{g^{(Z)(1)}_{j,k}(x_1), g^{(Z)(1)}_{\ell,m}(x_2)\right\} + O(n^{-1}).$$

That is, if $x \neq 0$ or $\pm\pi$, then

$$\mathrm{Cov}\left\{g^{(2)}_{j,k}(x), g^{(2)}_{\ell,m}(x)\right\} = n^{-1} W_n g^{(Z)}_{j,\ell}(x) g^{(Z)}_{m,k}(x) + o(n^{-1} W_n), \tag{6.15}$$

while if $x = 0$ or $\pm\pi$, then

$$\mathrm{Cov}\left\{g^{(2)}_{j,k}(x), g^{(2)}_{\ell,m}(x)\right\} = n^{-1} W_n \left\{g^{(Z)}_{j,\ell}(x) g^{(Z)}_{m,k}(x) + g^{(Z)}_{j,m}(x) g^{(Z)}_{\ell,k}(x)\right\} + o(n^{-1} W_n), \tag{6.16}$$

and if $x_1 \neq \pm x_2$, then

$$\mathrm{Cov}\left\{g^{(2)}_{j,k}(x_1), g^{(2)}_{\ell,m}(x_2)\right\} = o(n^{-1} W_n).$$

If $u_{j,k}(r)$ is unknown, the value c in (6.10) may be used to form a truncated estimate $C_{j,k}(r)$ of it as in (2.20) and (2.21). That is,

$$C_{j,k}(r) = \begin{cases} c & C^*_{j,k}(r) \leq c, \\ C^*_{j,k}(r) & \text{otherwise,} \end{cases}$$

where for $r \geq 0$

$$C^*_{j,k}(r) = (n - r)^{-1} \sum_{t=1}^{n-r} Y_j(t) Y_k(t+r)$$

and $C^*_{j,k}(-r) = C^*_{k,j}(r)$. The estimate $C_{j,k}(r)$ may then be substituted in (6.14) to give $g^{(3)}_{j,k}(x)$, say. As before, it may be shown that this substitution does not affect the asymptotic mean or variance of the estimate of $g_{j,k}(x)$.

6.3 Some Special Cases

There are cases in which the formula (6.13), giving $g^{(Z)}_{j,k}(x)$ in terms of $g_{j,k}(x)$, simplifies to a certain extent. One is where the series, or some subset of series, are either all observed or all missed at any particular time. That is, $P\{Y_j(t) = Y_k(t)\} = 1$, $j,k = 1, \ldots, h$. This might occur if some recording apparatus simultaneously recorded a number of series, and omissions were only caused by complete breakdown.

Suppose for instance that the first two series were recorded in this way. Then $p_j = p$ and $u_{j,k}(r) = u(r)$, for $j,k = 1,2$, and so

$$g^{(Z)}_{j,k}(x) = p^2 g_{j,k}(x) + \int_{-\pi}^{\pi} g_{j,k}(y)a(x-y)\,dy,$$

$j,k = 1,2$.

Thus each of these cross-spectral density functions is transformed in the same way, and the same sequence $u(r)$ or $C_{j,k}(r)$ is involved in the construction of $g^{(2)}_{j,k}(x)$ or $g^{(3)}_{j,k}(x)$, $j,k = 1,2$.

Another case in which (6.13) simplifies is where $Y_j(s)$ and $Y_k(t)$ are independent for each s and t. For then $u_{j,k}(r) = p_j\, p_k$ for each r, and so

$$g^{(Z)}_{j,k}(x) = p_j p_k\, g_{j,k}(x).$$

It follows that the estimate of $g_{j,k}(x)$ is simply a multiple of $g^{(Z)(1)}_{j,k}(x)$.

This type of deletion would evidently occur in many

situations, for example where completely separate apparatus was used to record different series. As another example, Brillinger (1969) performed a cross-spectral analysis of monthly sunspot numbers, Santa Fé rainfall, English rainfall and various European temperature figures. In the absence of indications to the contrary, it seems reasonable to assume that omissions in one series would be independent of those in any other.

A special case of the second example arises when one component of $\{\underline{X}(t)\}$ is always observed. For then one component of $\{\underline{Y}(t)\}$ is in effect nonrandom, and all other components are independent of it.

CHAPTER 7. A SIMULATION STUDY

The expressions which have been derived in Chapters 2, 3 and 4 for the variances of certain spectral estimates are all asymptotic. Thus, in general, they are not exact for any finite value of n, the length of the series being analysed. They do, however, give an arbitrarily close approximation for sufficiently large n. In order to assess the degree of approximation, a simulation study was carried out on an IBM 7094 computer at Imperial College.

The following procedure was used. First, 250 terms of a second-order autoregressive process were generated by

$$X(t) = 0.9X(t-1) - 0.8X(t-2) + U(t),$$

where $\{U(t)\}$ was a pseudo-random sample from the standard normal distribution. The series was started with $X(-1) = X(0) = 0$.

Then 250 terms of a renewal amplitude modulating process (Section 2.5), $\{Y(t)\}$, were generated, with $f_1 = 8/9$, $f_2 = 1/9$, $f_r = 0$ otherwise. This series was started with $Y(1) = 1$. The corresponding incomplete series $\{Z(t)\}$ was then obtained.

Next, 250 terms of a complementary renewal process (Section 2.5), $\{Y^*(t)\}$, were generated, with $f_8^* = f_9^* = f_{10}^* = 1/3$, $f_r^* = 0$ otherwise. This series was started with $Y^*(1) = 0$. Again, the amplitude modulated process

$\{z^*(t)\}$ was obtained.

It is clear that none of these series is stationary as constructed. The first 50 terms of each series was discarded, and the remainder treated as stationary. The relevant estimates for the incomplete series are those of Sections 2.3 and 2.4, the common asymptotic variance of which is given by (2.21).

The estimate $g_X^{(1)}(x)$ was then calculated from $X(t)$ for a set of values in $[0,\pi]$, and the estimates $g_X^{(2)}(x)$ and $g_X^{(3)}(x)$ were calculated from both the series $z(t)$ and $z^*(t)$, for the same values of x.

The weighting function used was the Parzen kernel (Parzen, 1961), for which

$$w(y;x) = \frac{3}{8\pi}\left[\frac{\sin\{\lambda_n(x-y)/4\}}{\sin\{(x-y)/4\}}\right]^4,$$

$$w_\gamma(x) = k(x/\lambda_n)\, e^{ix\gamma},$$

where

$$k(x) = \begin{cases} 1 - 6x^2 + 6|x|^3 & |x| < \frac{1}{2}, \\ 2(1-|x|)^3 & \frac{1}{2} \leqslant |x| < 1, \\ 0 & 1 \leqslant |x|. \end{cases}$$

and $W_n \sim 0.54\lambda_n$.

The quantity λ_n is known as the truncation point of the estimate, since $w_\gamma(x) = 0$ for $|x| \geqslant \lambda_n$, and so no covariance estimates of higher lag than λ_n are used. The choice of a weighting function and a truncation point has been widely discussed in the statistical literature; see for

example Jenkins and Watts (1968), Parzen (1961) or Jenkins (1961). The value chosen was $\lambda_{200} = 50$.

This procedure was repeated 100 times, giving a sample of 100 values for each of the five estimates, at each value of x. The mean and standard deviation of each sample was then calculated, and compared with the corresponding asymptotic value.

It was found that for both the renewal data and the complementary renewal data, the estimates $g_X^{(2)}(x)$ and $g_X^{(3)}(x)$ had virtually identical sample means and standard deviations. This is to be expected from the results of Section 2.4, for it may be seen there that the proof that $g_X^{(2)}(x)$ and $g_X^{(3)}(x)$ have the same asymptotic mean and variance shows incidently that their correlation is $1 + o(n^{-1} W_n)$. Since $g_X^{(3)}(x)$ is the more practical of the two estimates, the results for $g_X^{(2)}(x)$ are not given.

The sample and asymptotic means and standard deviations of $g_X^{(1)}(x)$, and of the estimate $g_X^{(3)}(x)$ based on the renewal data and the complementary renewal data, are given in Tables 1 to 3. The same information is given graphically in Figures 1 to 3.

The estimate $g_X^{(1)}(x)$ was calculated in order to provide some indication of how good the approximations are in the usual case of complete data. It may be seen that the approximations in the cases of missing data are comparable. This suggests that the results of Chapter 2 provide approximations

which are not appreciably worse than those normally used.

Of course, this does not imply anything about the results of Chapters 3 and 4. However, the similarity of the methods suggests that these also may give useful approximations in the finite sample case.

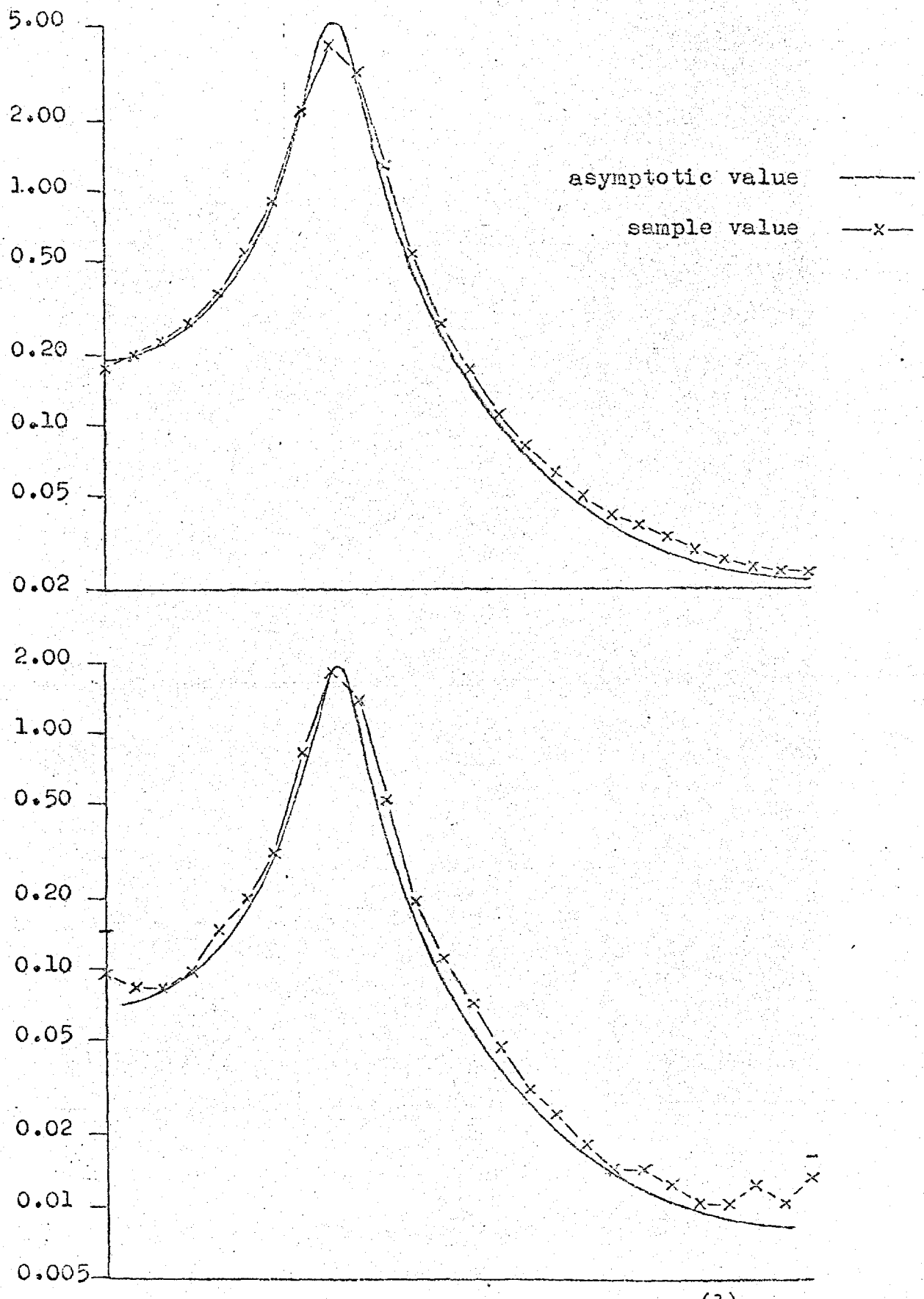


FIGURE 1. The mean and standard deviation of $g_X^{(1)}(x)$, the estimate based on a complete sample. The upper graph shows the mean, the lower graph the standard deviation.

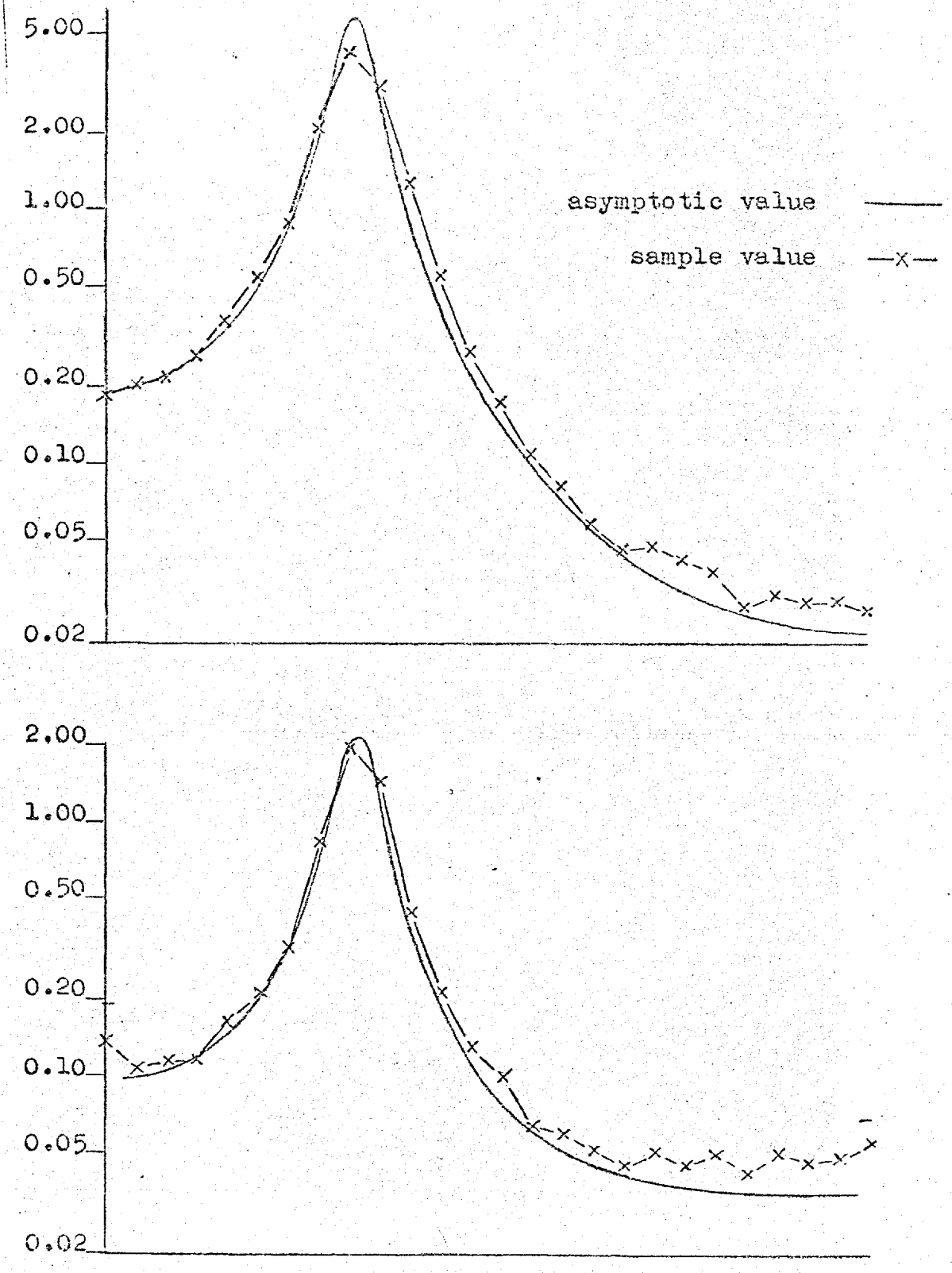


FIGURE 2. The mean and standard deviation of $g_X^{(3)}(x)$ calculated from the renewal-deleted data. The upper graph shows the mean, the lower graph the standard deviation.

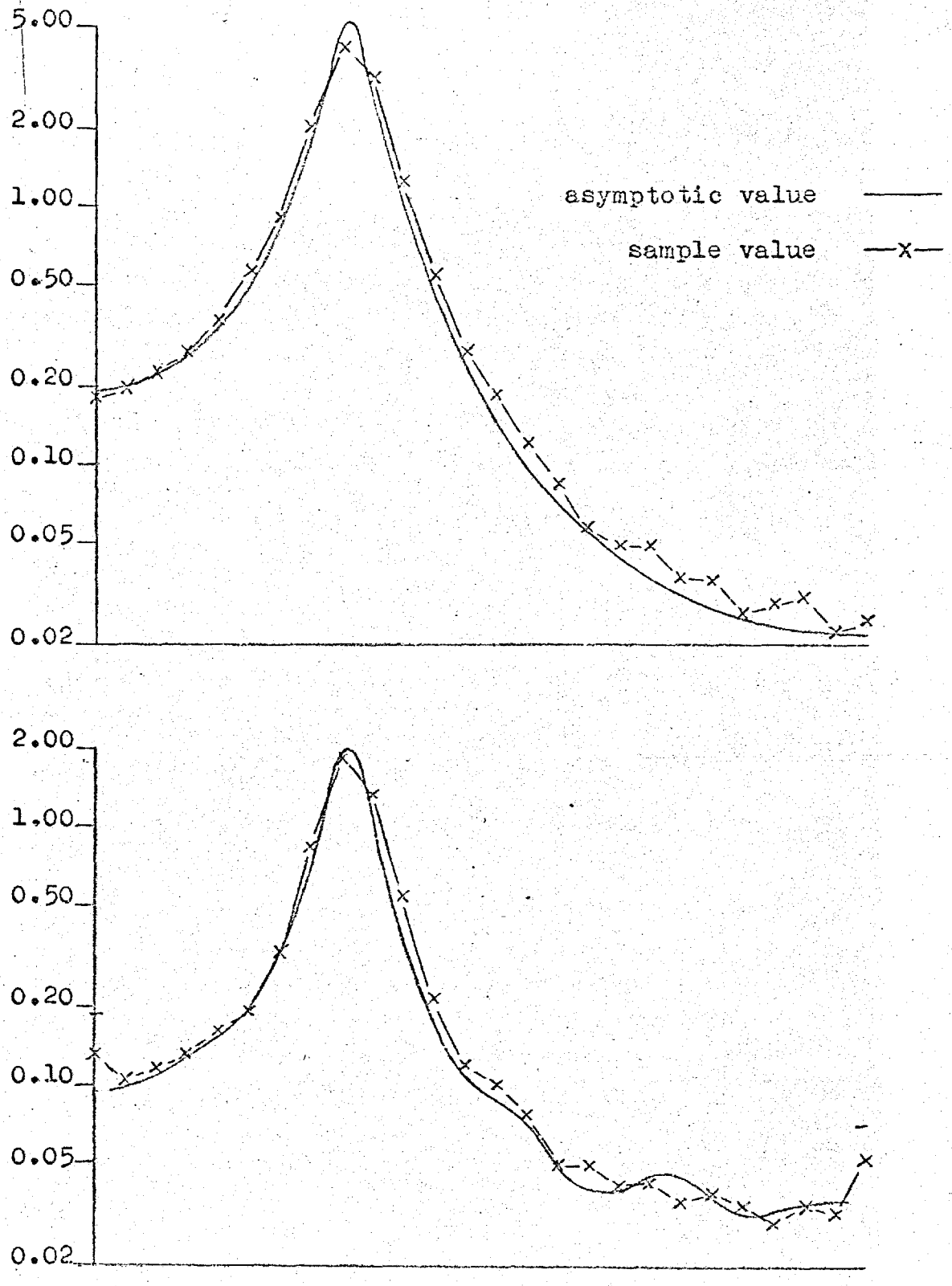


FIGURE 3. The mean and standard deviation of $g_X^{(3)}(x)$ calculated from the complementary renewal-deleted data. The upper graph shows the mean, the lower graph the standard deviation.

TABLE 1. The mean and standard deviation of $g_X^{(1)}(x)$, the estimate based on a complete sample

Frequency	Mean		Standard Deviation	
$(25x/\pi)$	Sample	Asymptotic	Sample	Asymptotic
0	0.179	0.196	0.096	0.144
1	0.201	0.203	0.084	0.074
2	0.229	0.223	0.083	0.082
3	0.277	0.264	0.098	0.097
4	0.367	0.340	0.148	0.125
5	0.552	0.492	0.197	0.181
6	0.897	0.844	0.314	0.310
7	2.131	1.874	0.825	0.688
8	4.110	4.871	1.807	1.788
9	3.143	3.146	1.373	1.154
10	1.291	1.031	0.510	0.378
11	0.540	0.448	0.197	0.164
12	0.269	0.241	0.110	0.089
13	0.174	0.149	0.072	0.055
14	0.112	0.101	0.047	0.037
15	0.084	0.074	0.031	0.027
16	0.063	0.057	0.024	0.021
17	0.050	0.045	0.018	0.017
18	0.042	0.038	0.014	0.014
19	0.038	0.032	0.014	0.012
20	0.034	0.029	0.012	0.010
21	0.030	0.026	0.010	0.010
22	0.027	0.024	0.010	0.009
23	0.025	0.023	0.012	0.008
24	0.024	0.022	0.010	0.008
25	0.024	0.022	0.013	0.016

TABLE 2. The mean and standard deviation of $g_X^{(3)}(x)$ calculated from the renewal deleted data

Frequency	Mean		Standard Deviation	
$(25x/\pi)$	Sample	Asymptotic	Sample	Asymptotic
0	0.186	0.196	0.135	0.190
1	0.203	0.203	0.108	0.097
2	0.222	0.223	0.113	0.105
3	0.267	0.264	0.115	0.120
4	0.370	0.340	0.162	0.148
5	0.543	0.492	0.213	0.205
6	0.873	0.844	0.321	0.335
7	2.098	1.874	0.841	0.714
8	4.123	4.871	1.912	1.815
9	3.151	3.146	1.419	1.176
10	1.278	1.031	0.438	0.400
11	0.550	0.448	0.222	0.187
12	0.277	0.241	0.130	0.112
13	0.178	0.149	0.099	0.078
14	0.112	0.101	0.064	0.061
15	0.083	0.074	0.060	0.052
16	0.058	0.057	0.051	0.046
17	0.046	0.045	0.045	0.042
18	0.048	0.038	0.051	0.039
19	0.043	0.032	0.045	0.038
20	0.038	0.029	0.050	0.036
21	0.028	0.026	0.042	0.036
22	0.031	0.024	0.050	0.035
23	0.029	0.023	0.046	0.035
24	0.029	0.022	0.048	0.034
25	0.027	0.022	0.055	0.069

TABLE 3. The mean and standard deviation of $g_X^{(3)}(x)$ calculated from the complementary renewal deleted data

Frequency	Mean		Standard Deviation	
$(25x/\pi)$	Sample	Asymptotic	Sample	Asymptotic
0	0.185	0.196	0.131	0.183
1	0.201	0.203	0.108	0.099
2	0.231	0.223	0.118	0.113
3	0.276	0.264	0.131	0.134
4	0.372	0.340	0.164	0.152
5	0.565	0.492	0.198	0.201
6	0.903	0.844	0.337	0.330
7	2.106	1.874	0.851	0.709
8	4.069	4.871	1.852	1.808
9	3.106	3.146	1.347	1.169
10	1.273	1.031	0.540	0.393
11	0.547	0.448	0.222	0.180
12	0.277	0.241	0.121	0.109
13	0.190	0.149	0.104	0.089
14	0.123	0.101	0.077	0.069
15	0.085	0.074	0.049	0.050
16	0.058	0.057	0.050	0.040
17	0.049	0.045	0.042	0.039
18	0.049	0.038	0.043	0.045
19	0.037	0.032	0.036	0.044
20	0.036	0.029	0.039	0.037
21	0.027	0.026	0.035	0.032
22	0.029	0.024	0.030	0.033
23	0.031	0.023	0.035	0.035
24	0.023	0.022	0.033	0.036
25	0.025	0.022	0.053	0.071

REFERENCES

BLOOMFIELD, P. (1970). Lower bounds for renewal sequences and p-functions. Z. Wahrscheinlichkeitstheorie (to appear).

BRILLINGER, D.R. & ROSENBLATT, M. (1967). Asymptotic theory of estimates of k-th order spectra. in Advanced Seminar on Spectral Analysis, ed. Harris. Wiley : New York.

BRILLINGER, D.R. (1969). A search for a relationship between monthly sunspot numbers and certain climatic series. Bull. Inst. Int. Statist., 43 (1), 293-306.

COOLEY, J.W., LEWIS, P.A.W. & WELCH, P.D. (1967). Historical notes on the Fast Fourier Transform. Trans. I.E.E.E. on Audio and Electroacoustics, AU-15, 76-84.

FELLER, W. (1968). An Introduction to Probability Theory and its Applications (3rd. ed.; Volume 1). New York : Wiley.

JENKINS, G.M. (1961). General considerations in the analysis of spectra. Technometrics, 3, 133-166.

JENKINS, G.M. & WATTS, D.G. (1968). Spectral Analysis and its Applications. San Francisco : Holden-Day.

JONES, R.H. (1962). Spectral analysis with regularly missed observations. Ann. Math. Statist., 33, 455-461.

NEAVE, H.R. (1970). Spectral analysis of a stationary time series using initially scarce data. Biometrika, 57, 111-122.

PARZEN, E. (1957) On consistent estimates of the spectrum of a stationary time series. Ann. Math. Statist., 28, 329-348.

PARZEN,E. (1961). Mathematical considerations in the estimation of spectra. Technometrics, 3, 167-190.

PARZEN,E. (1962). Spectral analysis of asymptotically stationary time series. Bull. Inst. Int. Statist., 39 (2), 1-18.

PARZEN,E. (1963). On spectral analysis with missing observations and amplitude modulation. Sankhyā A, 25, 383-392.

ROSENBLATT,M. (1959). Statistical analysis of stochastic processes with stationary residuals. in Probability and Statistics, ed. Grenander. Wiley : New York.

SCHEINOK,P.A. (1965). Spectral analysis with randomly missed observations : the binomial case. Ann. Math. Statist., 36, 971-977.