SOME ASPECTS OF LINEAR PREDICTION IN STATIONARY TIME SERIES

by

W. Kenneth Poole

University of North Carolina

Institute of Statistics Mimeo Series No. 566

This research was partially supported by the Mathematics Division of the Air Force Office of Scientific Research Contract No. AF-AFOSR-68-1415.

DEPARTMENT OF STATISTICS

University of North Carolina

Chapel Hill, N. C.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

The author extends a special thanks and appreciation to his advisor, Professor W. L. Smith, for the encouragement and advice received during this research.

He also wishes to thank the other committee members: Professor N. L. Johnson, Professor M. R. Leadbetter, Professor J. E. Grizzle, and Professor B. H. Kaplan for their genuine interest in the work.

Finally, he is particularly grateful to his wife, Cassandra for her patience and understanding and to Mrs. Bobbie Lou Wallick, Mrs. Ruby Monk, and Cassandra for their excellent typing of the manuscript.

ABSTRACT

Suppose X_t, $t = 0, \pm 1, \pm 2, \ldots$ is a second order stationary time series with zero mean and known autocorrelation function. The classical linear prediction problem deals with prediction of X_{t+h} from $X_t, X_{t-1}, \ldots, X_{t-n+1}$ for $h \geq 1$, where the predictors are restricted to be of the form $\sum_{i=0}^{n-1} a_i(n) X_{t-i}$ for finite n and limits in mean of such linear combinations for an infinite past. In practice it often happens that a sequence of numbers $a_0, a_1, \ldots$ exists such that $\sum_{i=0}^{\infty} a_i X_{t-i}$ is a mean convergent series which represents the best linear predictor of X_{t+h} based on an infinite past. When this is the case the a_i's are often used to predict from a large but finite past.

The problem of when such a coefficient sequence $\{a_i\}$ can be found, is the first one investigated in this research. It is related to the problem of determining when a non-orthogonal basis for a general Hilbert space forms a "Schauder basis" for that space.

The second problem deals with linear prediction when the mean of the process is unknown. Formulae are derived for the best linear predictor when the observed past is finite and the mean is a polynomial. If the observed past is infinite and the mean unknown no best linear prediction formulae exist so linear predictors that are "almost" best are constructed using a principle suggested by D. R. Cox.

SUMMARY AND REMARKS

In this thesis we have been concerned with several aspects of prediction theory in a stationary time series X_t, $t = 0, \pm 1, \pm 2, \ldots$. The existence of a best linear predictor for arbitrary stationary processes with known mean and correlation function was established independently by Kolmogorov and Weiner around 1940. This best predictor is, of course, a simple linear function of the observations for a finite observed past whereas if the observed past is infinite it is given as a limit-in-mean of a sequence of finite linear combinations of the observations. That is we predict X_{t+h} from X_t, $X_{t-1}, \ldots$ by a limit-in-mean of linear combinations of the form $\sum_{i=0}^{n-1} a_i(n)\, X_{t-i}$. In practice it often happens that the $a_i(n)$ can be chosen to be independent of n in this infinite case and the predictor is then given by a mean-convergent series $\sum_{i=0}^{\infty} a_i\, X_{t-i}$.

The second chapter of this work characterizes the time series whose best linear predictors from an infinite past are of the form $\sum_{i=0}^{\infty} a_i\, X_{t-i}$ for all $h \geq 1$. These processes are given the name <u>strictly linear</u> and form a class we call $\mathcal{S}$. $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$ are subclasses of $\mathcal{S}$ formed by restricting the coefficient sequence a_0, a_1, ... for h = 1. If the Hilbert space $\mathcal{H}$ generated by the X_t process has the property that every element Z in $\mathcal{H}$ may be represented in the form $\sum_{-\infty}^{+\infty} a_i\, X_i$ where the a_i's depend on Z then we say the X_t process is <u>completely linear</u> and all such processes form the class $\mathcal{C}$ (the terms strict and complete linearity are not meant to imply that $\mathcal{S}$ and $\mathcal{C}$ are subclasses of the linear processes of Grenander and Rosenblatt [4]). The condition that the X_t process be in $\mathcal{C}$ is closely related to though not equivalent to the condition that X_t, $t = 0, \pm 1, \pm 2, \ldots$ form a so called "Schauder basis" for $\mathcal{H}$. The elements of a Schauder basis are required to be linearly independent in the usual sense whereas we may have processes which are singular in the sense defined in chapter 1, page 6. If we require the X_t process to be nonsingular then the conditions are equivalent.

Conditions on the spectral distribution function which ensure strict linearity are found in chapter 2 and these are rather more general than previously known conditions.

The third chapter of this work is directed at the problem of finding the best linear predictor when the mean function is unknown, the correlation function is known, and the observed past is finite. Formulae are derived for the best linear predictor when the mean function is some polynomial. Properties of this predictor are discussed and examples given.

The last chapter is concerned with prediction from an infinite past when the mean function is unknown. As might be expected there is generally no "best" predictor in this case which is a linear function of the observations. The best "linear" predictor in this case is given as a limit-in-mean of a sequence of finite linear combinations of the past and generally involves the unknown mean function. With this in mind, an idea put forth by D. R. Cox is explored and shown to be useful in constructing good predictors when the infinite past is observed and the mean is an unknown constant. Efficient predictors, in a certain defined sense, are constructed using an extension of Cox's idea for the case of a constant mean. Best predictors, among those in a certain restricted class, are derived for processes with a polynomial mean function. These predictors are shown to be efficient for a useful family of processes.

Probably the most interesting problems to theoretical statisticians suggested by this work are those connected with the study of time series in $\mathcal{S}$. In light of past work done on autoregressive schemes of finite order it seems that stronger conditions on the spectral distribution function of a process in $\mathcal{S}$ should be deducible than those given in theorem 2.1. It also seems that necessary and sufficient conditions for $\mathcal{S}$-processes should be available if a spectral density is assumed. Theorem 2.2 gives the best conditions we have so far. Last of all, the connection between $\mathcal{S}$ and $\mathcal{C}$ needs further investigation. It is clear that $\mathcal{C} \subset \mathcal{S}$ but to what extent the converse relation holds still needs studying. Theorems 2.3 and 2.4 assure us that an $\mathcal{S}_2$ process with a bounded spectral density is in $\mathcal{C}$.

On the more practical side, the question of how good the predictors developed in chapter 4 are when truncated and applied to a large but finite past is one that needs investigation. Perhaps Tchebycheff-type bounds which depend only on the constants in the predictor, can be derived.

CHAPTER I

INTRODUCTION

A stochastic process in a family of random variables $\{X_t\}$ indexed by some set T. If the set T is the whole real line or some interval then X_t is a random function and if T is any countable subset of the real line X_t is called a random sequence. When the parameter t is interpreted as denoting time then the stochastic process is called a time series, with the adjective continuous or discrete being used to indicate whether it is a random function or a random sequence.

In this work we shall confine attention to complex-valued random sequences which may be represented by the model:

$$X_t = M_t + Y_t, \quad t \in T \qquad (1.1)$$

$$T = \{0, \pm 1, \pm 2, \ldots\}$$

where M_t in a nonrandom mean value function and Y_t in a fluctuation function. The Y_t process is assumed to be second order stationary with a mean function which is zero for all $t \in T$. This means

$$\text{a)} \quad E(Y_t) = 0, \quad t \in T$$

$$\text{b)} \quad E(Y_t \overline{Y_{t+\nu}}) = \sigma^2 \rho_{|\nu|} \qquad t, \nu \in T \qquad \sigma^2 < \infty \qquad (1.2)$$

where E denotes expectation and $\rho_{|\nu|}$ is the autocorrelation function of the process.

Of considerable interest in applied fields, as well as in the theory of stochastic processes, is the matter of predicting future values of a process after it has been observed at a number of time points. In particular, suppose we have observed X_t, $X_{t-1}, \ldots X_{t-n+1}$ for some integers t and n and wish to forecast the value of the process at t + h (i.e. predict X_{t+h}) utilizing the information supplied by the observed values, the forecast being optimum in some sense. Although our principle concern is with prediction it will develop that some of the results obtained do not depend on h being a positive integer and hence are valid for interpolating processes and in estimating past values of the process. Interpolation is useful when values of the observed sequence are missing for one reason or another.

Consider then as a predictor of X_{t+h} an arbitrary function of the past $f(X_t, \ldots X_{t-n+1})$. Since $E|X_t|^2 < \infty$, a reasonable criterion for optimality is to require that f be so chosen that

$$E|X_{t+h} - f(X_t, X_{t-1}, \ldots, X_{t-n+1})|^2 \tag{1.3}$$

is a minimum. We know immediately that the f minimizing (1.3) is given by the formula

$$f_{\min.}(X_t, \ldots, X_{t-n+1}) = E\{X_{t+h}|X_t, \ldots, X_{t-n+1}\}. \tag{1.4}$$

However, this presents a problem since we do not know the distribution of the vector $(X_{t+h}, X_t, X_{t-n+1})$. We see by this that we can not hope to solve the general problem without massive extra assumptions.

If f is suitably restricted then we may hope for a solution under only second order stationarity assumptions. We shall in particular require that f be a linear function of the past observations.

If M_t and $\rho_{|\nu|}$ are known functions then our problem becomes the classical prediction theory problem for discrete time series. Norbert Wiener and A. N. Kolmogorov, using somewhat different methods, independently solved this problem around 1940. Kolmogorov dealt only with the discrete case while Wiener treated both, giving more attention to the continuous case.

Since we will use some of the ideas and results of the classical theory a brief review will be given here, without proofs, for the discrete case. A more detailed review may be found in Grenander and Rosenblatt [4] for the discrete case and in the general case one may refer to E. A. Robinson [14].

Since M_t is assumed known we may take $E(X_t) = 0$ and consider the linear space $\mathcal{H}$ consisting of all the random variables X_t, $t \varepsilon T$, together with all finite linear combinations $\sum_{j=1}^{n} a_j X_{t_j}$ where $\{t_1, t_2, \ldots, t_n\} \subset T$ and all random variables which are limits-in-mean of sequences of such finite linear combinations, the norm being that defined in (1.6). The a_j's are complex constants. If for any two elements X, Y of $\mathcal{H}$ we define an inner product $\langle X,Y\rangle$ by

$$\langle X,Y\rangle = E(X\overline{Y}) \tag{1.5}$$

and a norm by

$$||X||^2 = \langle X,X\rangle \tag{1.6}$$

then $\mathcal{H}$ is a Hilbert space with norm $||\cdot||$.

For any sequence $\{X_t, X_{t-1}, \ldots, X_{t-n+1}\}$ there is a corresponding linear manifold M_n^t generated by all linear combinations of the elements of the sequence. In case n is finite the manifold is automatically closed and if $n = \infty$ we require that M_∞^t contain not only all finite linear combinations of elements of the sequence X_t, $X_{t-1}, \ldots$ but also all random variables that are limits in the mean of sequences of finite linear combinations. Then we have M_∞^t closed.

If we have observed X_t, $X_{t-1}, \ldots, X_{t-n+1}$ for $n < \infty$, our problem is the choosing of a_0, $a_1, \ldots, a_{n-1}$ to minimize

$$E|X_{t+h} - \sum_{i=0}^{n-1} a_i X_{t-i}|^2 \tag{1.7}$$

and this is precisely the same problem as choosing $Z_n \varepsilon M_n^t$ which minimizes

$$E|X_{t+h} - Z_n|^2. \tag{1.8}$$

From the theory of Hilbert spaces we know such a Z_n, Z_n^* say, exists and is given by the projection of X_{t+h} on M_n^t. This projection is uniquely determined by

$$E\{Z_n^* \overline{(Z_n^* - X_{t+h})}\} = 0 \tag{1.9}$$

which means the error of prediction is orthogonal to the predictor. It is in fact orthogonal to all $Z_n \varepsilon M_n^t$.

In case $n = \infty$, the problem of minimizing (1.7) relative to the a_i's is not equivalent to choosing $Z_\infty^* \varepsilon M_\infty^t$ which minimizes (1.8). The former problem may not have a solution whereas the latter one always does. Z_∞^* is chosen as the projection of X_{t+h} on M_∞^t. This projection may be obtained directly or as the limit-in-mean of the sequence Z_1^*, $Z_2^*, \ldots$. All solutions are unique in the sense that if Z_n^* and Z_n^{**} both minimize (1.8) then

$$E|Z_n^* - Z_n^{**}|^2 = 0. \tag{1.10}$$

If we put

$$\sigma^2_{n,h} = \min_{Z_n \varepsilon M_n^t} E|X_{t+h} - Z_n|^2 \tag{1.11}$$

then a rather obvious fact in that $\sigma^2_{n,h} \geq \sigma^2_{n+k,h}$ for $n > 1$, $k \geq 0$, $h > 0$, and a less obvious fact is that

$$\lim_{n\to\infty} \sigma^2_{n,h} = \sigma^2_{\infty,h}. \tag{1.12}$$

There are various methods of obtaining the projections of X_{t+h} on the M_n^t. Unfortunately, the most elegant theoretically is in general the least useful practically. If we suppose the X_t to be linearly independent in the usual sense, then M_n^t is spanned by n orthonormal elements which may be constructed from the X_t's by the Gram-Schmidt process of orthogonalization. Denote this orthonormal basis as $B_t, B_{t-1}, \ldots B_{t-n+1}$. The projection Z_n^* is then given by

$$Z_n^* = \sum_{i=0}^{n-1} B_{t-i}\, E\{X_{t+h}\, \overline{B_{t-i}}\} \tag{1.13}$$

and this is so for any n. Computationally, a better method for finite n is as follows:

Let $\underline{X}_n' = (X_t, X_{t-1}, \ldots, X_{t-n+1})$ and put

$$E\{\underline{X}_n\, \overline{\underline{X}}_n'\} = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \overline{\rho}_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \overline{\rho}_{n-1} & \overline{\rho}_{n-2} & & \cdots & 1 \end{pmatrix} = \sigma^2 \Sigma_n, \text{ say.}$$

Then

$$E\{X_{t+h}\,\underline{\overline{X}}_n'\} = \sigma^2(\rho_h, \rho_{h+1}, \ldots, \rho_{h+n-1}) = \sigma^2\underline{\rho}_n', \text{ say.}$$

Thus the vector $\underline{\hat{a}}_n' = (\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_{n-1})$, minimizing (1.7) is given by

(1.14) $$\underline{\hat{a}}_n' = \underline{\rho}_n' \Sigma_n^{-1}$$

and further

(1.15) $$\sigma^2_{n,h} = \sigma^2(1 - \underline{\rho}_n' \Sigma_n^{-1} \underline{\rho}_n) .$$

It is to be noted that if $X_t, X_{t-1}, \ldots, X_{t-n+1}$ are linearly dependent (i.e. Σ_n is singular) then X_{t+h} can be determined without error since a linear relation among the X_i's, $i = t, t-1, \ldots, t-n+1$, implies a linear relation among X_{t+h} and the X_i's, due to stationarity. This singular case will not be dealt with in this chapter. We shall call a process singular if prediction from an infinite observed past without error. All other processes will be called nonsingular.

In case $n = \infty$, $\underline{\hat{a}}_\infty'$ cannot generally be determined as when n is finite (i.e. by minimizing (1.7)). This is due to the fact that, although the best "linear" predictor exists (i.e. the projection on the closed linear manifold M_∞^t exists), it may not be expressible in the form $\sum_{i=0}^{\infty} a_i X_{t-i}$. This, in turn, is due to the fact that limits-in-mean of sequences of linear functions $\sum_{j=1}^{n} a_j X_{t_j}$, $\{t_1, t_2, \ldots, t_n\}$ T, may not themselves be linear functions. Chapter 2 is devoted entirely to this problem of best linear prediction based on an infinite past.

If the solution happens to be of the form $\sum_{i=0}^{\infty} a_i X_{t-i}$

then we can arrive at it as when n is finite, that is, we solve the system of equations

(1.17) $$\sum_{i=0}^{\infty} a_i \rho_{|i-j|} = \rho_{h+j} \qquad j = 0,1,2,\ldots$$

for the a_i's by any of several techniques. The solution to (1.17) will be unique if the process is not singular since

$$\sum_{i=0}^{\infty} a_i \rho_{|i-j|} = \rho_{h+j} \qquad j = 0,1,2,\ldots$$

and

$$\sum_{i=0}^{\infty} a_i^* \rho_{|i-j|} = \rho_{h+j} \qquad j = 0,1,2,\ldots$$

means that

$$\sum_{i=0}^{\infty} (a_i - a_i^*)\rho_{|i-j|} = 0.$$

This, in turn, means that

$$\sum_{j=0}^{\infty} \sum_{i=0}^{\infty} (a_j - a_j^*)(a_i - a_i^*)\rho_{|i-j|} = 0$$

which requires that (see lemma 2.1)

$$a_i = a_i^* \quad i = 0,1,2,\ldots\ .$$

On the other hand, if the optimum predictor is not of the form $\sum_{i=0}^{\infty} a_i X_{t-i}$ then (1.17) will not have a solution satisfying (1.16). In any case we can always get the required solution by solving the problem for the finite case and then taking the limit-in-mean.

If the parameters M_t and $\rho_{|\nu|}$ are not known then we have quite another problem. It has been suggested by some authors (Yaglom [22], page 99) that if both these functions are unknown we cannot hope for a solution with only second order stationarity assumptions. This seems clear enough since $\rho_{|\nu|}$ would necessarily have to be estimated, in some way, by some nonlinear function of the observations and immediately we are concerned with higher moments than second of the process if the mean square error in the criterion of optimality.

In chapters 3 and 4 we shall deal principally with the case when M_t is unknown but $\rho_{|\nu|}$ is known. Specifically, we will be concerned with (constructing a predictor, $\hat{X}_{t+h}$, of the value of the process at time t+h under the following conditions:)

$$C_1.\quad \hat{X}_{t+h} = \sum_{i=0}^{n-1} a_i X_{t-i}$$

$$C_2.\quad E(\hat{X}_{t+h}) = E(X_{t+h}),\ t \in T$$

$$C_3.\quad E|\hat{X}_{t+h} - X_{t+h}|^2 \text{ to be minimized} \tag{1.18}$$

$$C_4.\quad E\{X_t\} = M_t$$

where M_t is a polynomial of degree m.

The first requirement is essential if we are to work only with second order stationary processes. Any more complicated function would require further knowledge of the process if the mean square error is used to determine goodness. The condition C_2 is also essential since we are not assuming a known mean, otherwise the mean square error of prediction would depend on the unknown mean and the coefficients in the derived predictor would also depend on it. The third condition is set in the hope that it can be satisfied. Unfortunately, this will not always be the case. When (C_3) cannot be satisfied we will then look for predictors satisfying (C_1), (C_2), (C_4). These will often have mean square errors that are small as well as other nice properties. Condition (C_4) is for mathematical convenience, but at the same time has been used successfully in a great number of applications.

We are essentially dealing with a linear hypothesis model for the process where the mean of each observation lies on a polynomial mean function of degree m. The first work of this nature seems to have been done by Zadeh and Ragazinni [23] in 1949. They work exclusively with continuous time series and, employing methods similar to those of Wiener, they arrive at an integral equation similar to but more complicated than that of Wiener. They show by rather sophisticated Fourier methods that the equation can in principle be solved. Hájek [6] considers an approach similar to the Hilbert space one outlined above for the case of known mean. He deals with the general linear hypothesis model and allows the index set T to be completely general. In fact, the only real requirement placed on the process is that the correlation function not depend on the unknown parameters in the mean function.

Hájek shows by appealing to Hilbert space theory that if a linear function of the unknown parameters in the model is estimable in the usual sense then there exists a unique unbiased minimum variance estimator which is either a finite linear combination of variables X_t, $t \in T$, or is a limit-in-mean of a sequence of such linear functions. More generally, he shows that if a random variable Y has an estimable expectation then there is a unique unbiased minimum mean square error estimator of Y. Hájek's work gives a solution to our problem when n is finite but due to the generality with which he solves the problem his results are not in a useful form. His solutions appear as certain projections on closed linear manifolds of a particular Hilbert space. Our methods for the finite case are elementary and straightforward and they point out properties of the best predictor which are perhaps unexpected and not at all obvious from Hájek's work. This will be elaborated on later.

In the infinite case, it develops that no solution exists satisfying our conditions (C_1), (C_2), (C_3), (C_4), and whereas we can frequently find good predictors of the form desired, they will not be best. Hájek assures us of a best predictor but it, of course, will not be of the form $\sum_{i=0}^{\infty} a_i X_{t-i}$ that we require.

The only other work similar to ours seems to be that of D. R. Cox [3] who investigated linear predictors when the whole past of a process is known. Cox only considers the simplest case of a first order autoregressive scheme and predictors of the form

$$\hat{X}_{t+h} = (1 - \lambda) \sum_{i=0}^{\infty} \lambda^i X_{t-i}. \tag{1.19}$$

Indeed, it was this work of Cox which prompted the present investigation of arbitrary linear functions of the past for second order stationary processes. Cox's work is discussed and extended later in this thesis.

Chapter 2

PREDICTION FROM AN INFINITE PAST
WITH KNOWN MEAN FUNCTION

In this chapter we wish to look, in some detail, at a problem raised in chapter 1 concerning linear predictors from an infinite past when the mean function is known (i.e. M_t is a known function in (1.1)).

It was pointed out there that although there always exists a best "linear" predictor of X_{t+h} in the sense that it is the projection of X_{t+h} on M_{∞}^{t}, this projection may not in fact be of the form

$$\sum_{k=0}^{\infty} a_k X_{t-k} \tag{2.1}$$

for all $h \geq 1$. The following example will illustrate this point.

Suppose X_t is a stationary time series with the decomposition

$$X_t = Y_t^{(1)} + Y_t^{(2)} \tag{2.2}$$

where

a) $Y_t^{(1)}$ is a real first order autoregressive scheme with zero means, unit variance, and autocorrelation function

$$E(Y_t^{(1)} Y_{t+\tau}^{(1)}) = \rho^{|\tau|}, \ \tau = 0, \pm 1, \ldots, \ |\rho| < 1$$

b) $Y_t^{(2)}$ is a singular process independent of $Y_t^{(1)}$ and generated by taking an observation Y from a standard normal distribution and setting

$$Y_t^{(2)} = Y \quad t = 0, \pm 1, \ldots .$$

We will show that the best linear predictor of X_{t+h} is $\hat{Y}^{(1)}_{t+h} + \hat{Y}^{(2)}_{t+h}$ where $\hat{Y}^{(1)}_{t+h}$, $\hat{Y}^{(2)}_{t+h}$ are respectively the best linear predictors of $Y^{(1)}_{t+h}$ and $Y^{(2)}_{t+h}$.

It is well known that the best linear predictor of $Y^{(1)}_{t+h}$ based on $Y^{(1)}_{t}$, $Y^{(1)}_{t-1}, \ldots, Y^{(1)}_{t-n+1}$ is

$$\hat{Y}^{(1)}_{t+h} = \rho^h Y^{(1)}_t. \tag{2.3}$$

Consider then the variable

$$\hat{X}_{t+h} = \underset{k\to\infty}{\text{l.i.m}}\{\rho^h X_t + (1-\rho^h)\frac{1}{k}\sum_{i=0}^{k-1} X_{t-i}\}$$

$$= \underset{k\to\infty}{\text{l.i.m}}\sum_{i=0}^{k-1} b_i(k)\, X_{t-i}, \text{ say.} \tag{2.4}$$

As $\frac{1}{k}\sum_{i=0}^{k-1} Y^{(1)}_{t-i}$ is a mean consistent estimate of its expectation we have

$$\hat{X}_{t+h} = \rho^h X_t + (1-\rho^h)\underset{k\to\infty}{\text{l.i.m}}\frac{1}{k}\sum_{i=0}^{k-1} X_{t-i}$$

$$= \rho^h X_t + (1-\rho^h)\{Y^{(2)}_t + \underset{k\to\infty}{\text{l.i.m}}\sum_{i=0}^{k-1} Y^{(1)}_{t-i}/k\}$$

$$= \rho^h X_t + (1-\rho^h) Y^{(2)}_t \tag{2.5}$$

so that (2.4) does exist. Since the $\hat{X}_{t+h}$ defined above is in M^t_∞ (see chapter 1) we need only verify that $X_{t+h} - \hat{X}_{t+h}$ is orthogonal to M^t_∞ to show that $\hat{X}_{t+h}$ is the best "linear" predictor. From (2.5)

$$\hat{X}_{t+h} = \rho^h X_t + (1-\rho^h) Y_t^{(2)}$$

$$= \rho^h (Y_t^{(1)} + Y_t^{(2)}) + (1-\rho^h) Y_t^{(2)}$$

$$= \rho^h Y_t^{(1)} + Y_t^{(2)}$$

and we have

$$E\{(X_{t+h} - \hat{X}_{t+h}) X_{t-i}\} = E\{(Y_{t+h}^{(1)} - \hat{Y}_{t+h}^{(1)}) X_{t-i}\}$$

$$= E\{(Y_{t+h}^{(1)} - \hat{Y}_{t+h}^{(1)}) Y_{t-i}^{(1)}\}$$

$$+ E\{(Y_{t+h}^{(1)} - \hat{Y}_{t+h}^{(1)}) Y_{t-i}^{(2)}\}, \quad i = 0,1,2,\ldots$$

The first term on the right is zero since $\hat{Y}_{t+h}^{(1)}$ is the best linear predictor of $Y_{t+h}^{(1)}$ and the error of prediction is orthogonal to the $Y_{t-i}^{(1)}$'s. The second term is zero since the $Y_t^{(1)}$ and the $Y_t^{(2)}$ processes are independent. Hence we have that

$$\hat{X}_{t+h} = \rho^h Y_t^{(1)} + Y_t^{(2)}$$

$$= \hat{Y}_{t+h}^{(1)} + \hat{Y}_{t+h}^{(2)} \tag{2.6}$$

is the best linear predictor based on an infinite past. Suppose now that

$$\hat{X}_{t+h} = \sum_{i=0}^{\infty} b_i X_{t-i}.$$

As

$$\hat{X}_{t+h} = \lim_{k\to\infty} \sum_{i=0}^{k-1} b_i(k) X_{t-i}$$

we must have from lemma 2.1, page 39, that

$$\lim_{k\to\infty} b_i(k) = b_i, \quad i = 0,1,2,\ldots \; .$$

But

$$\lim_{k\to\infty} b_i(k) = \lim_{k\to\infty} \frac{(1 - \rho^h)}{k}$$

$$= 0$$

if $i \geq 1$ and

$$\lim_{k\to\infty} b_0(k) = \rho^h$$

and hence

$$\hat{X}_{t+h} = \rho^h X_t .$$

But from (2.5) we have

$$\hat{X}_{t+h} = \rho^h X_t + (1 - \rho^h) Y_t^{(2)}$$

and this is a contradiction since

$$E\{(1 - \rho^h) Y_t^{(2)}\}^2 = (1 - \rho^h)^2$$

$$\neq 0.$$

Several authors (Yaglom [22], Wold [21]) have noted this difficulty of best linear prediction based on an infinite past but no one seems to have directed attention specifically at this problem. It will be our goal in this section to find conditions on the process which will ensure that the best linear predictor is expressable as (2.1). We find it convenient to work with the spectral representation of the process and its correlation function, so these concepts are introduced along with some results which will be helpful in our investigation. Proofs of the results will not be given as they may be found in standard texts on stationary time series (e.g. Yaglom [22], Wold[21], Grenander and Rosenblatt [4], Rosenblatt [15]).

Suppose X_t is a weakly stationary time series with covariance function $\sigma^2\rho_\tau$, then

$$\rho_\tau = \int_{-\pi}^{+\pi} e^{i\tau y}\, dF(y) \tag{2.8}$$

where F(y) is a distribution function on $(-\pi, +\pi)$. Conversely, for every distribution function on $(-\pi, +\pi)$ there exists a function defined by the integral on the right hand side of (2.8) which is the correlation function of a stationary time series. Hence a study of F(y) is equivalent to a study of ρ_τ. (2.8) is called the spectral representation of ρ_τ and F(y) is the spectral distribution function of the process. There is also a spectral representation of the process itself. This is

$$X_t = \int_{-\pi}^{+\pi} e^{ity}\, dZ(y) \tag{2.9}$$

where Z(y) is a continuous time stochastic process on the interval $(-\pi, +\pi)$ with the properties:

a) $E\{Z(y)\} = 0, \quad y \varepsilon (-\pi, +\pi)$

b) $E\{(Z(y_1) - Z(y_2))\overline{(Z(y_3) - Z(y_4))}\} = 0$

for $-\pi \leq y_1 \leq y_2 < y_3 \leq y_4 \leq +\pi$.

c) $E|Z(y_1) - Z(y_2)|^2 = \sigma^2\{F(y_2) - F(y_1)\}$

for $-\pi \leq y_1 < y_2 \leq +\pi$, where $\sigma^2 = E|X_t|^2$.

In other words, the $Z(y)$ process is one of zero mean function and orthogonal increments.

d) If $g(\lambda)$ is any function for which

$$\int_{-\pi}^{+\pi} |g(\lambda)|^2 \, dF(\lambda) < \infty \tag{2.10}$$

then the random Stieltjes integral

$$\int_{-\pi}^{+\pi} g(\lambda) dZ(\lambda)$$

is well defined and has the properties:

i) If $g(\lambda)$ and $h(\lambda)$ are two such functions then

$$\int_{\alpha_1}^{\alpha_2} (ag(\lambda) + bh(\lambda)) dZ(\lambda) = a\int_{\alpha_1}^{\alpha_2} g(\lambda) dZ(\lambda) + b\int_{\alpha_1}^{\alpha_2} h(\lambda) dZ(\lambda).$$

for $(\alpha_1, \alpha_2) \subset (-\pi, +\pi)$.

ii) $\underset{n\to\infty}{\text{l.i.m}} \int_{\alpha_1}^{\alpha_2} g_n(\lambda) dZ(\lambda) = \int_{\alpha_1}^{\alpha_2} g(\lambda) dZ(\lambda)$

if and only if

$$\lim_{n\to\infty} \int_{\alpha_1}^{\alpha_2} |g_n(\lambda) - g(\lambda)|^2 \, dF(\lambda) = 0$$

iii) $$E\left\{\int_{-\pi}^{+\pi} g(\lambda)dZ(\lambda) \overline{\int_{-\pi}^{+\pi} h(\lambda)dZ(\lambda)}\right\} = \sigma^2 \int_{-\pi}^{+\pi} g(\lambda)\overline{h(\lambda)}dF(\lambda).$$

These results concerning the spectral representation of the X_t process may be found in [15], chapter VII.

The relation (2.9) establishes an isomorphism between the Hilbert space $\mathcal{H}$, defined in chapter one, and the Hilbert space $\mathcal{H}_F$ of all functions on the interval $(-\pi, +\pi)$ which are square integrable with respect to the measure $F(y)$ in (2.8). The isomorphism is such that X_t corresponds to e^{ity}. In this section we will refer to the two Hilbert spaces as $\mathcal{H}$ and $\mathcal{H}_F$ with norms $||\cdot||$ and $||\cdot||_F$ respectively. The closed linear manifolds in $\mathcal{H}_F$ corresponding to M_n^t in $\mathcal{H}$ will be denoted by ${}_n^tM$. The best linear predictor of X_{t+h} will be of the form (2.1) if and only if its isomorph in $\mathcal{H}_F$ is of the form

(2.11) $$e^{ity} \sum_{k=0}^{\infty} a_k \, e^{-iky}.$$

We will need the following definitions.

<u>Definition 1</u>

Suppose X_t, $t = 0, \pm1, \pm2, \ldots$ is a weakly stationary time series with zero mean which satisfies

(2.12) $$\sum_{i=0}^{\infty} a_i \, X_{t-i} = \eta_t \qquad t = 0, \pm1, \pm2, \ldots$$

where

i) $a_0 \neq 0$

ii) $E(\eta_t) = 0$

iii) $E(\eta_t \overline{\eta}_\tau) = \sigma_\eta^2 \, \delta_{t,\tau}$

where $\delta_{t,\tau}$ is the Kronecker delta function.

iv) $E(\eta_t \overline{X}_\tau) = 0, \quad t > \tau$

then X_t is called a generalized autoregressive scheme. It should be emphasized that the series on the left of (2.12) is assumed convergent in the mean and also that σ_η^2 is not assumed nonzero.

<u>Definition 2</u>

Suppose for the stationary time series X_t, the best linear predictor of X_{t+h} based on X_t, $X_{t-1}, \ldots$ is of the form (2.1) for all $h \geq 1$. Then the time series is said to be <u>strictly linear</u>. This class of time series will be denoted by $\mathscr{L}$.

<u>Definition 3</u>

Let $\mathscr{H}$ be the Hilbert space generated by the process in definition 2. If $Z \epsilon \mathscr{H}$ implies that

(2.13) $$Z = \sum_{-\infty}^{+\infty} a_i X_i$$

then the process is said to be <u>completely linear</u>. This class we designate by $\mathfrak{C}$.

<u>Theorem 2.1</u>

The stationary time series X_t with zero mean and unit variance is a member of $\mathscr{L}$ if and only if X_t is a generalized autoregressive scheme. If the X_t process is nonsingular and $\sum_{k=0}^{\infty} a_k e^{-ik\lambda}$ has a countable number of zeros then the continuous part of the spectral distribution function of X_t is absolutely continuous with spectral density

$$F_2'(\lambda) = \frac{\sigma_\eta^2}{2\pi\left|\sum_{k=0}^{\infty} a_k e^{-ik\lambda}\right|^2} \tag{2.14}$$

where the a_k's are the coefficients of the scheme.

Proof:

Suppose X_t is a time series in $\mathcal{S}$. Consider for an arbitrary t the problem of predicting one step ahead. The best linear predictor is of the form

$$\hat{X}_{t+1} = \sum_{i=0}^{\infty} b_{1,i} X_{t-i} \tag{2.15}$$

due to the strict linearity of X_t. The coefficients $b_{1,i}$ do not depend on t since the process is stationary.

Define

$$\begin{aligned} \eta_{t+1} &= X_{t+1} - \sum_{i=0}^{\infty} b_{1,i} X_{t-i} \\ &= \sum_{i=0}^{\infty} c_i X_{t+1-i}, \text{ say.} \end{aligned} \tag{2.16}$$

The right hand side of (2.16) is convergent in the mean since best linear predictors must be. We also have

$$c_0 = 1$$

and since

$$E(X_t) = 0$$

for all t, then

$$E(\eta_t) = 0.$$

Consider $\eta_{\tau+1}$ for $\tau \geq t$. As $\eta_{\tau+1}$ is the error of prediction of $X_{\tau+1}$ from X_τ, $X_{\tau-1}, \ldots$ then $\eta_{\tau+1}$ is orthogonal to the linear manifold M_∞^τ. In particular,

$$E(\eta_{\tau+1} \overline{X}_t) = 0.$$

For $t < \tau$ we have

$$\eta_{t+1} \in M_\infty^{t+1}$$

while

$$M_\infty^{t+1} \subset M_\infty^\tau$$

and so

$$E(\eta_\tau \overline{\eta}_t) = \sigma_\eta^2 \delta_{\tau,t}.$$

Then the X_t process is generalized autoregressive.

Suppose now that the series is generalized autoregressive and we wish to predict X_{t+h} from an observed past X_t, $X_{t-1}, \ldots$. Let $\hat{X}_{t+i}$ be the best linear predictor of X_{t+i} based on X_t, $X_{t-1}, \ldots$. We have that

$$X_{t+1} = -\sum_{i=0}^{\infty} \frac{a_i}{a_0} X_{t-i} + \eta_{t+1}/a_0$$

$$= \sum_{i=0}^{\infty} b_{1,i} X_{t-i} + \overset{*}{\eta}_{t+1}, \text{ say,}$$

since $a_0 \neq 0$ where $\overset{*}{\eta}_{t+1}$ enjoys the same orthogonality properties as η_{t+1}. We now consider the variables $\hat{X}_{t+i}$ defined inductively as follows and show that they are the best linear predictors:

$$\hat{X}_{t+1} = \sum_{i=0}^{\infty} b_{1,i} X_{t-i}$$

$$\hat{X}_{t+2} = b_{1,0} \hat{X}_{t+1} + \sum_{i=1}^{\infty} b_{1,i} X_{t+1-i}$$

$$= \sum_{i=0}^{\infty} b_{2,i} X_{t-i}, \text{ say.}$$

$$\vdots$$

$$\hat{X}_{t+h} = \sum_{j=0}^{h-2} b_{1,j} \hat{X}_{t+h-j-1} + \sum_{j=h-1}^{\infty} b_{1,j} X_{t+h-j-1}$$

$$= \sum_{j=0}^{\infty} b_{h,j} X_{t-j}, \text{ say.}$$

It is clear that $\hat{X}_{t+h} \varepsilon M_{\infty}^{t}$ since $\hat{X}_{t+i}$, $i = 1,2,\ldots, h-1$ are all in M_{∞}^{t} and

$$\sum_{j=h-1}^{\infty} b_{1,j} X_{t+h-j-1} \varepsilon M_{\infty}^{t}.$$

Also $\hat{X}_{t+h}$ is of the form (2.1) so we need only show that $\hat{X}_{t+h} - X_{t+h}$ is orthogonal to M_{∞}^{t} and we do this by showing that

$$\hat{X}_{t+h} = X_{t+h} + \sum_{i=0}^{h-1} d_{h,i} \eta^{*}_{t+h-i} \tag{2.17}$$

where the $d_{h,i}$'s are constants. (2.17) is certainly true for $h = 1$ since

$$\hat{X}_{t+1} = \sum_{i=0}^{\infty} b_{1,i} X_{t-i}$$

$$= X_{t+1} - \eta^{*}_{t+1}.$$

Assume that

$$\hat{X}_{t+j} = X_{t+j} + \sum_{i=0}^{j-1} d_{j,i} \eta^{*}_{t+j-i}$$

for $j < k$. Then

$$\hat{X}_{t+k} = \sum_{i=0}^{k-2} b_{1,i} \hat{X}_{t+k-i-1} + \sum_{i=k-1}^{\infty} b_{1,i} X_{t+k-i-1}$$

$$= \sum_{i=0}^{k-2} b_{1,i}(X_{t+k-i-1} + \sum_{\ell=0}^{k-i-2} d_{k-i-1,\ell} \eta^*_{t+k-i-\ell-1})$$

$$+ \sum_{i=k-1}^{\infty} b_{1,i} X_{t+k-i-1}$$

$$= \sum_{i=0}^{\infty} b_{1,i} X_{t+k-i-1} + \sum_{i=0}^{k-2} \sum_{\ell=0}^{k-i-2} b_{1,i} d_{k-i-1,\ell} \eta^*_{t+k-i-\ell-1}.$$

The double sum is obviously a linear function of the η^*_τ's and since $t+k-1-(i+\ell)$ takes the maximum value $t+k-1$ and the minimum value $t+1$ on the range of summation we have

$$\hat{X}_{t+k} = X_{t+k} - \eta^*_{t+k} + \sum_{i=0}^{k-2} g_{k,i} \eta^*_{t+k-i-1}$$

$$= X_{t+k} + \sum_{i=0}^{k-1} d_{k,i} \eta^*_{t+k-i}$$

and $\hat{X}_{t+h}$ is expressible as claimed. This means that

$$\hat{X}_{t+h} - X_{t+h} = \sum_{i=0}^{h-1} d_{h,i} \eta^*_{t+h-i} \tag{2.18}$$

and since η^*_{t+i}, $i = 1,2,\ldots, h$ is orthogonal to M^t_∞ so is the left hand side of (2.18). Then $\hat{X}_{t+h}$ as defined above is the best linear predictor of X_{t+h} based on X_t, $X_{t-1}, \ldots$.

To prove the result about the spectral distribution function we first note that since

$$\sum_{i=0}^{\infty} a_i X_{t-i} = \eta_t$$

for $t = 0, \pm 1, \pm 2, \ldots$ then we must have

$$\int_{-\pi}^{+\pi} \sum_{k=0}^{\infty} a_k \, e^{i(t-k)\lambda} \, dZ_x(\lambda) = \int_{-\pi}^{+\pi} e^{it\lambda} \, dZ_\eta(\lambda)$$

in accord with (2.9). Letting $A(e^{-i\lambda})$ be the function $\sum_{k=0}^{\infty} a_k \, e^{-ik\lambda}$ we have

$$\int_{-\pi}^{+\pi} e^{it\lambda} \, A(e^{-i\lambda}) dZ_x(\lambda) = \int_{-\pi}^{+\pi} e^{it\lambda} \, dZ_\eta(\lambda). \tag{2.19}$$

Now consider the function $g(\lambda)$ on $(-\pi, +\pi)$ defined by

$$\begin{aligned} g(\lambda) &= 1, \quad -\pi \leq \lambda \leq \mu \\ &= 0, \quad \mu < \lambda \leq +\pi. \end{aligned}$$

Since $g(\lambda)$ is bounded with a square integrable derivative (almost everywhere), there exists a sequence of trigonometric polynomials, $g_n(\lambda)$, such that $g_n(\lambda)$ converges uniformly and absolutely to $g(\lambda)$. This implies that

$$\lim_{n\to\infty} \int_{-\pi}^{+\pi} |g_n(\lambda) - g(\lambda)|^2 |A(e^{-i\lambda})|^2 \, dF(\lambda) = 0 \tag{2.20}$$

and

$$\lim_{n\to\infty} \frac{\sigma_\eta^2}{2\pi} \int_{-\pi}^{+\pi} |g_n(\lambda) - g(\lambda)|^2 \, d\lambda = 0$$

where $F(\lambda)$ is the spectral distribution function of the X_t process and $\sigma_\eta^2(\lambda+\pi)/2\pi$ is the spectral distribution function of the η_t process. In view of this we have

$$\int_{-\pi}^{+\pi} \sum_{-n}^{n} b_t \, e^{it\lambda} \, A(e^{-i\lambda}) dZ_x(\lambda) = \int_{-\pi}^{+\pi} \sum_{-n}^{+n} b_t \, e^{it\lambda} \, dZ_\eta(\lambda) \tag{2.21}$$

where

$$g_n(\lambda) = \sum_{-n}^{+n} b_t \, e^{it\lambda}.$$

Hence

$$\underset{n\to\infty}{\text{l.i.m}} \int_{-\pi}^{+\pi} g_n(\lambda) A(e^{-i\lambda}) \, dZ_x(\lambda) = \underset{n\to\infty}{\text{l.i.m}} \int_{-\pi}^{+\pi} g_n(\lambda) \, dZ_\eta(\lambda)$$

and

$$\int_{-\pi}^{\mu} A(e^{-i\lambda}) \, dZ_x(\lambda) = \int_{-\pi}^{\mu} dZ_\eta(\lambda)$$

for $-\pi \leq \mu \leq +\pi$. Using property d.(iii) on page 17 we have

$$E\left|\int_{-\pi}^{\mu} A(e^{-i\lambda}) \, dZ_x(\lambda)\right|^2 = E\left|\int_{-\pi}^{\mu} dZ_\eta(\lambda)\right|^2$$

as

$$\int_{-\pi}^{\mu} |A(e^{-i\lambda})|^2 \, dF(\lambda) = \frac{\sigma_\eta^2(\mu + \pi)}{2\pi} . \tag{2.22}$$

If the X_t process is nonsingular then $\sigma_\eta^2 > 0$ since this is the mean square error of prediction one step ahead. Let $F_1(\lambda)$ and $F_2(\lambda)$ be the discrete and continuous parts of $F(\lambda)$ respectively. We first note that the left hand side of (2.22) must be continuous in μ since the right hand side is. But

$$\int_{-\pi}^{\mu} |A(e^{-i\lambda})|^2 \, dF(\lambda) = \int_{-\pi}^{\mu} |A(e^{-i\lambda})|^2 \, dF_1(\lambda) + \int_{-\pi}^{\mu} |A(e^{-i\lambda})|^2 \, dF_2(\lambda)$$

and so the points of increase of $F_1(\lambda)$, if any, must coincide with the zeroes of $A(e^{-i\lambda})$, otherwise the integral would not be continuous.

This means

$$\int_{-\pi}^{\mu} |A(e^{-i\lambda})|^2 \, dF_2(\lambda) = \sigma_\eta^2(\mu + \pi)/2\pi$$

and the measure on the left hand side must be absolutely continuous with respect to Lebesgue measure since the right hand side is. Then

$$\int_{-\pi}^{\mu} |A(e^{-i\lambda})|^2 \, F_2'(\lambda) d\lambda = \int_{-\pi}^{\mu} \frac{\sigma_\eta^2}{2\pi} \, d\lambda$$

for $-\pi \leq \mu \leq +\pi$ and hence

$$F_2'(\lambda) = \sigma_\eta^2/2\pi|A(e^{-i\lambda})|^2.$$

The remainder of the results in this chapter deal chiefly with conditions on the spectral distribution function of a process which ensure that the process is in some defined subclass of $\mathscr{A}$. We define the following subclasses:

If X_t is a process of the class $\mathscr{A}$ and the coefficient sequence associated with X_t satisfies

$$\sum_{k=0}^{\infty} |a_k|^2 < \infty$$

then X_t is said to be in $\mathscr{A}_1$. If the sequence is such that

$$\sum_{k=0}^{\infty} |a_k| < \infty$$

then X_t is in $\mathscr{A}_2$ and finally, if

$$|a_k| \leq M\rho^k$$

for some $0 \leq \rho < 1$, then X_t is a member of $\mathscr{A}_3$. We obviously have the relation

$$\mathcal{S}_3 \subset \mathcal{S}_2 \subset \mathcal{S}_1.$$

Theorem 2.2

Suppose X_t is a stationary time series with a spectral density function then X_t is a member of $\mathcal{S}_1$ if and only if

$$\int_{-\pi}^{+\pi} \frac{1}{f(\lambda)}\, d\lambda < \infty. \tag{2.23}$$

Proof:

Assume (2.23) is true. Since

$$\int_{-\pi}^{+\pi} \log f(\lambda) d\lambda < \int_{-\pi}^{+\pi} f(\lambda) d\lambda$$

we see that we must have

$$\int_{-\pi}^{+\pi} \log \frac{1}{f(\lambda)}\, d\lambda > -\infty. \tag{2.24}$$

Put $h(\lambda) = 1/f(\lambda)$. As $h(\lambda)$ is integrable and the integral of $\log h(\lambda)$ is not $-\infty$ we may appeal to the important theorem of Riesz and Fejér (see [5], page 20 and [4], Page 75) and get

$$\begin{aligned} h(\lambda) &= 2\pi \Big|\sum_{k=0}^{\infty} b_k\, e^{-ik\lambda}\Big|^2 \\ &= 2\pi |B(e^{-i\lambda})|^2, \text{ say,} \end{aligned} \tag{2.25}$$

where $B(z)$ satisfies the following conditions:

a) $B(z)$ is analytic and has no zeroes in $|z| < 1$.

b) $\sum_{k=0}^{\infty} |b_k|^2 < \infty$.

Hence

$$f(\lambda) = 1/h(\lambda)$$

$$= 1/2\pi|B(e^{-i\lambda})|^2.$$

In order to show that this representation implies that the process is a member of $\mathscr{S}_1$ we need to introduce the following class of functions. A function G(z) of the complex variable z is said to be in the class H_2 if

i) G(z) is analytic in $|z| < 1$

and ii) $\int_{-\pi}^{+\pi} |G(re^{i\theta})|^2 \, d\theta$ is bounded for all $r < 1$.

In the appendix of [4] it is proved that if G(z) is in H_2 then

$$\lim_{r\to 1-} \int_{-\pi}^{+\pi} |G(re^{i\theta}) - G(e^{i\theta})|^2 \, d\theta = 0. \tag{2.26}$$

Consider the variables

$$\eta_t = \sum_{k=0}^{\infty} b_k X_{t-k}$$

$t = 0, \pm 1, \pm 2, \ldots$. We can, of course, assume that $b_0 \neq 0$. Since $E(X_{t-k}) = 0$ for all t and k then $E(\eta_t) = 0$ for all t. For arbitrary t and τ we have

$$E(\eta_t \overline{\eta}_\tau) = \int_{-\pi}^{+\pi} \sum_{k=0}^{\infty} b_k e^{i(t-k)\lambda} \overline{\sum_{k=0}^{\infty} b_k e^{i(\tau-k)\lambda}} f(\lambda) d\lambda$$

$$= \int_{-\pi}^{+\pi} e^{i(t-\tau)\lambda} |\sum_{k=0}^{\infty} b_k e^{-ik\lambda}|^2 f(\lambda) d\lambda$$

$$= \frac{1}{2\pi} \int_{-\pi}^{+\pi} e^{i(t-\tau)\lambda} d\lambda$$

$$= \delta_{t,\tau}$$

and the η_t's are orthogonal. Consider next for $t < \tau$

$$E(\eta_\tau \overline{X}_t) = \int_{-\pi}^{+\pi} e^{-it\lambda} \sum_{k=0}^{\infty} b_k \, e^{i(\tau-k)\lambda} \, f(\lambda) d\lambda$$

$$= \frac{1}{2\pi} \int_{-\pi}^{+\pi} \frac{e^{i(\tau-t)\lambda}}{\sum_{k=0}^{\infty} \overline{b}_k \, e^{ik\lambda}} \, d\lambda$$

$$= \frac{1}{2\pi} \int_{-\pi}^{+\pi} K_\ell(e^{i\lambda}) d\lambda, \text{ say,}$$

where $\tau - t = \ell$. Now since $B(z)$ has no zeroes in $|z| < 1$ then $K_\ell(z)/z$ is analytic there for $\ell = 1,2,\ldots$, and since

$$\frac{1}{2\pi} \int_{-\pi}^{+\pi} |K_\ell(e^{i\lambda})|^2 \, d\lambda = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \frac{d\lambda}{|\sum_{k=0}^{\infty} b_k \, \overline{e}^{ik\lambda}|^2}$$

$$= 1$$

then $K_\ell(z)/z$ is a function in H_2 (a theorem by G. H. Hardy in [9] states that

$$\frac{1}{r^2} \int_{-\pi}^{+\pi} |K_\ell(re^{i\lambda})|^2 \, d\lambda$$

is a nondecreasing function of r). But for $r < 1$ we have

$$\int_{-\pi}^{+\pi} K_\ell(re^{i\lambda}) d\lambda = 0$$

due to the analyticity of $K_\ell(z)/z$ for $|z| < 1$. Then

$$\lim_{r \to 1-} \int_{-\pi}^{+\pi} K_\ell(re^{i\lambda}) d\lambda = 0$$

and in view of

$$\left| \int_{-\pi}^{+\pi} K_\ell(re^{i\lambda}) d\lambda - \int_{-\pi}^{+\pi} K_\ell(e^{i\lambda}) d\lambda \right| \le \int_{-\pi}^{+\pi} |K_\ell(re^{i\lambda}) - K_\ell(e^{i\lambda})| d\lambda$$

we see that (2.26) implies that

$$\lim_{r\to 1-} \int_{-\pi}^{+\pi} |K_\ell(re^{i\lambda}) - K_\ell(e^{i\lambda})|\, d\lambda = 0$$

and consequently

$$\lim_{r\to 1-} \int_{-\pi}^{+\pi} K_\ell(re^{i\lambda})\, d\lambda = \int_{-\pi}^{+\pi} K_\ell(e^{i\lambda})\, d\lambda$$
$$= 0.$$

Hence $E(\eta_\tau, \overline{X}_t) = 0$ for $\tau > t$ and this means that X_t is a member of $\mathcal{S}_1$.

Suppose now that X_t is a process in $\mathcal{S}_1$. This means that the spectral density function has the form

$$f(\lambda) = \frac{1}{2\pi |\sum_{k=0}^{\infty} a_k e^{-ik\lambda}|^2}.$$

We note that

$$\int_{-\pi}^{+\pi} |\sum_{k=0}^{\infty} a_k r^k e^{-ik\lambda}|^2 \, d\lambda = (\sum_{k=0}^{\infty} |a_k|^2 r^{2k}) 2\pi$$

and so

$$\lim_{r\to 1-} \int_{-\pi}^{+\pi} |\sum_{k=0}^{\infty} a_k r^k e^{-ik\lambda}|^2 \, d\lambda = (\sum_{k=0}^{\infty} |a_k|^2) 2\pi$$
$$< \infty.$$

This means that $\sum_{k=0}^{\infty} a_k z^k$ is a function in H_2 and as a result

$$\lim_{r\to 1-} \int_{-\pi}^{+\pi} |\sum_{k=0}^{\infty} a_k r^k e^{-ik\lambda} - \sum_{k=0}^{\infty} a_k e^{-ik\lambda}|^2 \, d\lambda = 0.$$

We conclude then that

$$\lim_{r\to 1-} \int_{-\pi}^{+\pi} \left|\sum_{k=0}^{\infty} a_k \, r^k \, e^{-ik\lambda}\right|^2 d\lambda = \int_{-\pi}^{+\pi} \left|\sum_{k=0}^{\infty} a_k \, e^{-ik\lambda}\right|^2 d\lambda$$

$$= \frac{1}{2\pi} \int_{-\pi}^{+\pi} \frac{1}{f(\lambda)} \, d\lambda$$

$$= \left(\sum_{k=0}^{\infty} |a_k|^2\right) 2\pi < \infty .$$

<u>Theorem 2.3</u>

Suppose X_t is a process in $\mathscr{S}_2$ which has a spectral density, $f(\lambda)$. Then there exists ε such that

$$f(\lambda) \geq \varepsilon > 0$$

on $(-\pi, +\pi)$.

Proof:

If X_t is in $\mathscr{S}_2$ then

$$f(\lambda) = \frac{1}{2\pi \left|\sum_{k=0}^{\infty} a_k \, e^{-ik\lambda}\right|^2} .$$

As

$$\left|\sum_{k=0}^{\infty} a_k \, e^{-ik\lambda}\right|^2 \leq \left(\sum_{k=0}^{\infty} |a_k|\right)^2$$

then

$$f(\lambda) \geq \frac{1}{\left(\sum_{k=0}^{\infty} |a_k|\right)^2}$$

$$> 0.$$

Putting

$$\varepsilon = 1/(\sum_{k=0}^{\infty}|a_k|)^2$$

the result follows.

The next theorem goes in the converse direction and concerns completely linear spaces.

Theorem 2.4

Suppose X_t is a stationary time series with a spectral density which satisfies

$$(2.27) \qquad 0 < \varepsilon \le f \le M < \infty \qquad \text{a.e on } (-\pi, +\pi)$$

then X_t is a process in $\mathcal{C}$.

Proof:

Let Z be an element of $\mathcal{H}$ and $g(\lambda)$ its isomorph in $\mathcal{H}_f$. Since Z is a limit-in-mean of a sequence of finite linear combinations of X_t, $t = 0, \pm 1, \pm 2, \ldots$ we may write

$$\lim_{n\to\infty}\left|\left|\sum_{-n}^{+n} b_k(n)X_k - Z\right|\right| = 0$$

or equivalently

$$\lim_{n\to\infty}\left|\left|\sum_{-n}^{n} b_k(n)e^{ik\lambda} - g(\lambda)\right|\right|_f = 0.$$

Further,

$$\int_{-\pi}^{+\pi}|g(y)|^2\, f(y)dy \ge \varepsilon\int_{-\pi}^{+\pi}|g(y)|^2\, dy$$

due to (2.27) so that

$$\int_{-\pi}^{+\pi}|g(y)|^2\, dy < \infty .$$

By Fourier theory g(y) has a series representation

$$g(y) \sim \sum_{-\infty}^{+\infty} a_K e^{iKy}$$

which means

$$\lim_{n\to\infty} \int_{-\pi}^{+\pi} |g(y) - \sum_{-n}^{+n} a_K e^{iKy}|^2 \, dy = 0.$$

Then by (2.27) we have

$$\lim_{n\to\infty} \int_{-\pi}^{+\pi} |g(y) - \sum_{-n}^{+n} a_K e^{iKy}|^2 f(y)dy = 0$$

or

$$\lim_{n\to\infty} E|Z - \sum_{-n}^{+n} a_K X_K|^2 = 0.$$

The conditions of this theorem seem a bit too demanding, particularly the requirement that f(y) be bounded above. As it has been only the smallness of f(y) that has been important in previous theorems in this chapter it is felt that weaker conditions should be available to ensure us that a process is in $\mathcal{C}$.

Since (2.27) implies that the X_t process is nonsingular we have actually proved that X_t, t = 0, ± 1, ± 2, ... forms a Schauder basis for the Hilbert space $\mathcal{H}$.

Theorem 2.5

The stationary time series X_t is a nonsingular process in $\mathcal{S}_3$ if and only if the spectral distribution function, $F(\lambda)$, is absolutely continuous with density

$$f(\lambda) = |C(e^{-i\lambda})|^2 \tag{2.28}$$

where $C(z)$ is analytic and nonzero in $|z| < r$ for some $r > 1$.

Proof:

Suppose X_t is a nonsingular process in $\mathcal{S}_3$. By theorem 2.1 the continuous part of $F(\lambda)$ is absolutely continuous and its density is

$$F_2'(\lambda) = \frac{1}{2\pi\left|\sum_{k=0}^{\infty} a_k e^{-ik\lambda}\right|^2} .$$

As $|a_k| < M\rho^k$ for some $0 \leq \rho < 1$ then $A(z) = \sum_{k=0}^{\infty} a_k z^k$ is analytic inside the circle $|z| < \frac{1}{\rho}$. This means that $A(e^{-i\lambda})$ cannot have zeroes on the unit circle for if there were a zero we would have

$$A(e^{-i\lambda}) = (e^{-i\lambda} - e^{-i\lambda_j})^{\ell}\, h(\lambda)$$

for some $-\pi \leq \lambda_j \leq +\pi$ and some positive integer ℓ. $h(\lambda)$ is bounded away from zero in a neighborhood of λ_j and also bounded above. Then

$$F_2'(\lambda) = \frac{1}{2\pi\left|e^{-i\lambda} - e^{-i\lambda_j}\right|^{2\ell}\left|h(\lambda)\right|^2}$$

in a neighborhood of λ_j so that $F_2'(\lambda)$ tends to ∞ and is of such an order in a neighborhood of λ_j that $F_2'(\lambda)$ cannot be integrable. Hence $A(z)$ cannot have zeroes on $|z| = 1$ and $F(\lambda)$ can have no jumps (recalled from the proof of theorem 2.1 that jumps in $F(\lambda)$ must coincide with zeroes of $A(e^{-i\lambda})$). We conclude then that the discrete part of $F(\lambda)$ is missing and hence

$$F'(\lambda) = \frac{1}{2\pi|A(e^{-i\lambda})|^2}.$$

Since $A(z)$ is analytic in $|z| < \frac{1}{\rho}$ then for each $r < \frac{1}{\rho}$ $A(z)$ can have only a finite number of zeroes in $|z| < r$. As $A(z)$ has no zeroes on $|z| = 1$ then there is a $\delta > 0$ such that $A(z)$ has no zeroes in the annulus $1 - \delta \leq |z| \leq 1 + \delta$. Hence all the zeroes of $A(z)$ which are inside the unit circle are inside $|z| < 1 - \delta$. Let these zeroes be $z_1, z_2, \ldots, z_p$ and put

$$B(z) = A(z) \prod_{i=1}^{p} \frac{(z\bar{z}_p - 1)}{(z - z_p)} .$$

$B(z)$ is an analytic function inside $|z| < \frac{1}{\rho}$ and has no zeroes inside $|z| < 1 + \delta$. Also

$$|B(e^{-i\lambda})| = |A(e^{-i\lambda})|$$

since the product of p terms has modulus one on the unit circle. We have then that

$$F'(\lambda) = \frac{1}{2\pi|B(e^{-i\lambda})|^2} .$$

But now $1/B(z)$ is analytic inside $|z| < 1 + \delta$ and has the expansion

$$\frac{1}{B(z)} = \sum_{k=0}^{\infty} C_k\, z^k$$

$$= C^*(z), \text{ say,}$$

where $C^*(z)$ has no zeroes inside $|z| < 1 + \delta$. Then

$$F'(\lambda) = \frac{1}{2\pi}|C^*(e^{-i\lambda})|^2$$

$$= |C(e^{-i\lambda})|^2$$

where C(z) has the properties claimed.

Suppose now that $F(\lambda)$ is absolutely continuous with the representation

$$F'(\lambda) = |C(e^{-i\lambda})|^2$$

$$= \frac{1}{2\pi}|C^*(e^{-i\lambda})|^2, \text{ say,}$$

where C(z) is analytic and nonzero in $|z| < r$, $r > 1$. Then

$$\frac{1}{C^*(z)} = \sum_{k=0}^{\infty} b_k z^k$$

$$= B(z) \text{ in } |z| < r$$

where B(z) has the same properties as C(z). From the proof of theorem 2.2 this representation of $F'(\lambda)$ implies that the X_t process is a member of $\mathcal{S}$. The fact that X_t cannot be singular follows from noting that if X_t were singular we would have

$$E|\eta_t|^2 = 0$$

where

$$\eta_t = \sum_{k=0}^{\infty} b_k X_{t-k}.$$

But

$$E|\eta_t|^2 = \frac{1}{2\pi}\int_{-\pi}^{+\pi}|\sum_{k=0}^{\infty} b_k\, e^{-ik\lambda}|^2\, f(\lambda)d\lambda$$

$$= 1.$$

Since $B(z)$ is analytic within and on every circle $|z| < r' < r$ we have by Cauchy's inequality

$$|b_k| \leq M_{r'}/(r')^k$$

where,

$$M_{r'} = \max_{|z|=r'} |B(z)|.$$

Take $r' = \frac{1 + r}{2}$ and put $\rho = 1/r'$. Then $0 \leq \rho < 1$ and we have $|b_k| \leq M_\rho\, \rho^k$.

Theorems 2.2, 2.3, and 2.5 are results which give some consequences of putting restrictions on the coefficients associated with a process in $\mathcal{A}$. The restrictions on these coefficients are also restrictions on the coefficients of the best linear predictor. This is apparent from the formulae relating the coefficients of the scheme and the coefficients of the predictors. Suppose X_t is in $\mathcal{A}$, then following the notation in the proof of theorem 2.1 we have

$$\sum_{k=0}^{\infty} a_k\, X_{t-k} = \eta_t$$

and

$$\hat{X}_{t+1} = -\sum_{k=0}^{\infty} \frac{a_{k+1}}{a_0} X_{t-k}$$

$$= \sum_{k=0}^{\infty} b_k X_{t-k}.$$

If $b_{h,k}$ is the coefficient of X_{t-k} in $\hat{X}_{t+h}$ we have from the proof of theorem 2.1 that

$$b_{1,k} = b_k$$

$$b_{2,k} = b_{1,k} b_0 + b_{k+1}$$

$$b_{3,k} = b_{2,k} b_0 + b_{1,k} b_1 + b_{k+2}$$

$$\vdots$$

$$b_{h,k} = b_{h-1,k} b_0 + b_{h-2,k} b_1 + \ldots + b_{1,k} b_{h-2} + b_{k+h-1}.$$

It is clear that the restrictions placed on the a_k sequence are true if and only if the same restrictions are true for the sequence $b_{h,1}$, $b_{h,2}, \ldots$ for arbitrary $h \geq 1$.

If only a finite number of the constants in a generalized autoregressive scheme are nonzero then the process is autoregressive in the usual sense. It is well known that for such processes the correlation function ρ_t satisfies the condition

$$|\rho_t| < M \rho^t \tag{2.29}$$

where $0 \leq \rho < 1$ and hence the correlation between distant variables tends to zero exponentially fast. This, in a manner of speaking, expresses an independence between X_t and the distant past of the process.

Another condition that implies a type of independence of X_t and the remote past is that of regularity of the process. A regular process is one for which

(2.30) $$\lim_{h\to\infty} ||\hat{X}_{t+h} - X_{t+h}|| = \sigma$$

where $\hat{X}_{t+h}$ is the best linear predictor based on X_t, $X_{t-1}, \ldots$ and σ^2 is the variance of the process. (2.30) is equivalent to

(2.31) $$\lim_{h\to\infty} ||\hat{X}_{t+h}|| = 0$$

which more clearly indicates an independence of X_{t+h} and the remote past.

A more restrictive condition on the process is the complete regularity of Y. A. Rozanov. This is defined as follows:

Let M_{∞}^{-t} and M_0^{∞} be the closed linear manifolds generated by X_{-t}, $X_{-t-1}, \ldots$ and X_0, $X_1, \ldots$ respectively. Put

(2.32) $$\rho(t) = \sup_{Z\epsilon M_{\infty}^{-t},\, Y\epsilon M_0^{\infty}} |\text{Corr.}(Z,Y)|.$$

The X_t process is completely regular if

$$\lim_{t\to\infty} \rho(t) = 0.$$

As regular process and some processes in $\mathscr{A}$ have a type of independence in common one is led to ask if there is any useful relation between these two types of processes. The following example shows that completely regular (and hence regular) processes are not necessarily in $\mathscr{A}$. Let Y_t be a process with the properties:

i) $E(Y_t) = 0$, all t

ii) $E(Y_t \, \overline{Y}_{t+\tau}) = \delta_{t,\tau}$, all t,τ.

Form the following variables:

$$X_t = \frac{1}{\sqrt{m}} \sum_{i=0}^{m-1} Y_{t-i}. \tag{2.33}$$

Then the X_t process has the properties:

i) $E(X_t) = 0$, all t

ii) $E(X_t \, \overline{X}_{t+\tau}) = \frac{m - |\tau|}{m}$, $\tau = 0, \pm 1, \pm 2, \ldots, \pm m-1$

$= 0$, $|\tau| \geq m$.

A simple calculation gives the spectral density function as

$$\begin{aligned} f(\lambda) &= \frac{1}{2\pi m} \frac{|1 - e^{im\lambda}|^2}{|1 - e^{i\lambda}|^2} \\ &= \frac{1}{2\pi m} \frac{\sin^2 (\frac{m\lambda}{2})}{\sin^2(\lambda/2)} . \end{aligned} \tag{2.34}$$

The X_t process is evidently completely regular since $E(X_t \, \overline{X}_{t+\tau}) = 0$ as soon as $|\tau| \geq m$ and therefore the manifolds M_{∞}^{-t} and M_0^{∞} are orthogonal if $t \geq m$. In order to show that the process is not in $\mathcal{S}$ we need the following two lemmas.

<u>Lemma 2.1</u>

If X_t is a nonsingular stationary time series

and $\tilde{X}_{t,n} = \sum_{i=0}^{n-1} a_i(n) X_{t-i}$

is such that $\lim_{n\to\infty} a_i(n) = c_i$ $\quad i = 0, 1, 2, \ldots$

and $\underset{n\to\infty}{1\cdot i\cdot m}\, \tilde{X}_{t,n} = \tilde{X}_{t,\infty}$

where

$$\tilde{X}_{t,\infty} = \sum_{i=0}^{\infty} a_i X_{t-i}$$

then

(2.35) $$c_i = a_i \qquad i = 0, 1, 2, \ldots .$$

Proof: We know

$$\lim_{n\to\infty} \left|\left|\sum_{i=0}^{n-1} b_i(n) X_{t-i}\right|\right| = 0$$

where

$$b_i(n) = a_i(n) - a_i \qquad i = 0, 1, 2, \ldots, .$$

Assume (2.35) is not true and let be the smallest integer for which

(2.36) $$c_K \neq a_K.$$

For $n > K > 0$ we have

$$\left|\ \left|\left|\sum_{i=0}^{K-1} b_i(n) X_{t-i}\right|\right| - \left|\left|\sum_{i=K}^{n-1} b_i(n) X_{t-i}\right|\right|\ \right| \leq \left|\left|\sum_{i=0}^{n-1} b_i(n) X_{t-i}\right|\right|$$

so that the left hand side has limit zero and since by assumption

$$\lim_{n\to\infty} \left|\left|\sum_{i=0}^{K-1} b_i(n) X_{t-i}\right|\right| = 0$$

we have

$$\lim_{n\to\infty} \left|\left|\sum_{i=K}^{n-1} b_i(n) X_{t-i}\right|\right| = 0.$$

For some integer $N > K$, $b_K(n) \neq 0$ for $n \geq N$ due to (2.36).

Hence

$$\left|\left|\sum_{i=K}^{n-1} b_i(n) X_{t-i}\right|\right| = |b_K(n)|\ \left|\left|\sum_{i=K}^{n-1} b_i^*(n) X_{t-i}\right|\right|$$

for $n \geq N$ where

$$b_i^*(n) = b_i(n)/b_K(n) \qquad i = K, K+1, \ldots, n-1.$$

Then

$$\lim_{n\to\infty} \left|\left|\sum_{i=K}^{n-1} b_i^*(n) X_{t-i}\right|\right| = \lim_{n\to\infty} \left|\left|\sum_{i=0}^{n-K-1} b_{i+K}^*(n) X_{(t-K)-i}\right|\right| = 0.$$

But this means that the mean square error of the best linear predictor of $X_{t-\kappa}$ from $X_{t-\kappa-1}$, $X_{t-\kappa-2}$, ... is zero, contradicting the hypothesis of nonsingularity. Hence (2.35) holds.

<u>Lemma 2.2</u>

If X_t is a stationary time series with $\mathrm{Var}(X_t) = \sigma^2 > 0$ and $\sum_{i=0}^{\infty} a_i X_{t-i}$ is convergent in the mean then

$$\lim_{k\to\infty} |a_k| = 0. \tag{2.37}$$

Proof:

If $\sum_{i=0}^{\infty} a_i X_{t-i}$ is convergent in the mean we have

$$\lim_{m,n\to\infty} \left|\left| \sum_{i=0}^{m} a_i X_{t-i} - \sum_{i=0}^{n} a_i X_{t-i} \right|\right| = 0,$$

that is, the sequence of partial sums is a Cauchy sequence. In particular

$$\begin{aligned} \lim_{n\to\infty} \left|\left| \sum_{i=0}^{n} a_i X_{t-i} - \sum_{i=0}^{n-1} a_i X_{t-i} \right|\right| &= \lim_{n\to\infty} || a_n X_{t-n} || \\ &= \lim_{n\to\infty} \sigma |a_n| \\ &= 0. \end{aligned}$$

As $\sigma > 0$ the result follows.

If we take m = 2 and h = 1 in our example, then the best linear predictor based on a finite past is (see [12])

$$\hat{X}_{t+1,n} = \frac{1}{n+1} \sum_{i=0}^{n-1} (-1)^i (n - i) X_{t-i}.$$

$$= \sum_{i=0}^{n-1} a_i(n) X_{t-i}, \text{ say.}$$

Hence

$$\lim_{n\to\infty} a_i(n) = (-1)^i$$

and by Lemma 2.1 the coefficients of the best linear predictor based on an infinite past must be

$$a_i = (-1)^i$$

if the process is in $\mathscr{A}$. But

$$\lim_{i\to\infty} |a_i| = 1$$

and this means that the X_t process cannot be in $\mathscr{A}$.

On the other hand, if we consider the singular process constructed as follows we see that strictly linear processes are not necessarily regular. Let X be an observation from a distribution G(x) and set

$$X_t = X, \quad t = 0, \pm 1, \pm 2, \ldots$$

then the X_t process is a stationary time series with spectral distribution function

$$F(y) = 0, \quad -\pi \leq y < 0$$

$$= 1, \quad 0 \leq y \leq +\pi.$$

Clearly the best linear predictor based on an infinite past is

$$\hat{X}_{t+h} = X_t, \quad h \geq 1$$

and the process is strictly linear. However, since

$$\rho_{|\tau|} = 1, \quad \tau = 0, \pm 1, \ldots$$

the process cannot be regular.

Summary of Chapter 2

In this section we have been concerned with the problem of determining when the best linear predictor of X_{t+h}, based on an infinite past, is a bona fide linear function of X_t, $X_{t-1}, \ldots$. We have assumed a zero mean and known correlation function for this investigation and have given processes with this property the name strictly linear. Theorem 2.1 characterizes this class of processes as those which satisfy a certain difference equation. Theorems 2.2, 2.3, and 2.5 give conditions which the spectral distribution function of the process must satisfy if certain restrictions are placed on the coefficients of the best linear prediction formulae. Theorem 2.4 shows that processes with a spectral density bounded away from zero and above must be in the class $\mathcal{C}$ of completely linear time series.

Regularity and complete regularity of time series is reviewed and examples are given which show that the class of regular processes neither contains nor is contained in the class $\mathcal{S}$ of strictly linear processes.

CHAPTER 3

PREDICTION FROM A FINITE PAST WHEN THE MEAN IS UNKNOWN

In this section we are primarily concerned with solving the problem of predicting X_{t+h} from a knowledge of X_t, $X_{t-1},\ldots,X_{t-n+1}$ under conditions (1.18) when n is finite. We shall obtain explicit expressions for the coefficients in the predictor and note some nice properties of the predictor. The basic results in this section are embodied in the following theorem. We assume the process real valued.

Theorem 3.1

If X_t, $X_{t-1},\ldots,$ X_{t-n+1} are observed from a nonsingular process with model (1.1) where M_t is a polynomial of degree m and $m < n$, then the unique best linear unbiased predictor of X_{t+h} is $\underline{\tilde{a}}'\ \underline{X}$ where

$$\underline{\tilde{a}} = \Sigma^{-1}\underline{\rho} - \Sigma^{-1}M(M'\Sigma^{-1}M)^{-1}(M'\Sigma^{-1}\underline{\rho} - \underline{h}). \tag{3.1}$$

The mean square error of prediction is

$$\sigma^2\{1 - \underline{\rho}'\Sigma^{-1}\underline{\rho} + (M'\Sigma^{-1}\underline{\rho} - \underline{h})'(M'\Sigma^{-1}M)^{-1}(M'\Sigma^{-1}\underline{\rho} - \underline{h})\} \tag{3.2}$$

where (certain subscripts have been suppressed for simplicity)

$$\underline{\rho}' = (\rho_h,\ \rho_{h+1},\ldots\ \rho_{h+n-1})$$

$$\underline{\tilde{a}}' = (\tilde{a}_0,\ \tilde{a}_1,\ldots\ \tilde{a}_{n-1})$$

$$\underline{X}' = (X_t, X_{t-1}, \ldots X_{t-n+1})$$

$$\underset{(m+1)\times n}{M'} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 2 & \cdots & n-1 \\ 0 & 1 & 2^2 & \cdots & (n-1)^2 \\ \vdots & & & & \\ 0 & 1 & 2^m & \cdots & (n-1)^m \end{pmatrix}$$

$$\underset{n\times n}{\Sigma} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \vdots & & & & \vdots \\ \rho_{n-1} & \rho_{n-2} & & \cdots & 1 \end{pmatrix}$$

$$\underline{h}' = (1, -h, h^2, \ldots, (-h)^m) .$$

This predictor has the following properties:

P_1. The predictor depends on t only through the observations, i.e., $\underline{\tilde{a}}'$ is independent of t.

P_2. The predictor is unbiased even if it is constructed under the assumption that the degree of M_t is m when in fact it is $m' < m$.

P_3. Under some conditions the predictor tends, in the mean, to the best linear predictor for known mean function as $n \longrightarrow \infty$.

P_4. Under the same conditions as in P_3, the mean square error of prediction is asymptotically the same whether we correctly guess m or guess too high.

Proof:

Conditions (C_1) and (C_2) above require that

$$E\{\hat{X}_{t+h}\} = \sum_{i=0}^{n-1} a_i \; E\{X_{t-i}\}$$

$$= \sum_{i=0}^{n-1} a_i \; M_{t-i}$$

$$= M_{t+h}$$

$$= E\{X_{t+h}\}, \quad t \in T.$$

Now

$$M_{t+h} = M_t + h\, M_t^{(1)} + \frac{h^2}{2!} M_t^{(2)} + \ldots + \frac{h^m}{m!} M_t^{(m)}$$

while

$$M_{t-i} = M_t - i\, M_t^{(1)} + \frac{i^2}{2!} M_t^{(2)} - \ldots \frac{(-i)^m}{m} M_t^{(m)}$$

where

$$M_t^{(j)} = d^j \, M_t / dt^j .$$

Hence we require that

$$\sum_{i=0}^{n-1} a_i \sum_{j=0}^{m} \frac{(-i)^j \, M_t^{(j)}}{j!} = \sum_{j=0}^{m} \frac{M_t^{(j)}}{j!} \sum_{i=0}^{n-1} (-i)^j \; a_i$$

$$= \sum_{j=0}^{m} \frac{h^j}{j!} M_t^{(j)}, \quad t \in T$$

and this is equivalent to requiring that the following m+1 conditions be satisfied.

$$\sum_{i=0}^{n-1} a_i = 1$$

$$\sum_{i=0}^{n-1} i\, a_i = -h$$

$$\vdots$$

(3.3) $$\sum_{i=0}^{n-1} i^m a_i = (-h)^m.$$

Hence to satisfy (C_2) and (C_3) we must minimize (1.7) subject to (3.3). In view of (C_2) we may write

$$E\{X_{t+h} - \sum_{i=0}^{n-1} a_i X_{t-i}\}^2 = E\{(X_{t+h} - M_{t+h}) - \sum_{i=0}^{n-1} a_i (X_{t-i} - M_{t-i})\}^2$$

$$= E\{Y_{t+h} - \sum_{i=0}^{n-1} a_i Y_{t-i}\}^2$$

(3.4) $$= E\{Y_{t+h}^2 - 2 \sum_{i=0}^{n-1} a_i Y_{t+h} Y_{t-i} + (\sum_{i=0}^{n-1} a_i Y_{t-i})^2\}$$

$$= \sigma^2\{1 - 2\, \underline{a}'\, \underline{\rho} + \underline{a}'\, \Sigma\, \underline{a}\}.$$

The restrictions (3.3) may be written as

(3.5) $$M'\, \underline{a} = \underline{h}$$

so we are faced with the minimization of

(3.6) $$F(\underline{a},\lambda) = 1 - 2\underline{a}'\underline{\rho} + \underline{a}'\Sigma\underline{a} + 2\underline{\lambda}'(M'\underline{a} - \underline{h})$$

where $\underline{\lambda}' = (\lambda_0, \lambda_1, \ldots, \lambda_m)$ is a vector of Lagrangian multipliers. Denote by $\partial F/\partial \underline{x}$ the column vector of partial derivatives of F relative to the components of $\underline{x}$. Then we have

$$\frac{\partial F}{\partial \underline{a}} = -2\underline{\rho} + 2\Sigma\underline{a} + 2M\underline{\lambda}$$

$$\frac{\partial F}{\partial \underline{\lambda}} = 2(M'\underline{a} - \underline{h}).$$

Equating these partials to zero then gives the equations

(3.7) $$\underline{\tilde{a}} = \Sigma^{-1}(\underline{\rho} - M\underline{\lambda})$$

(3.8) $$M'\underline{\tilde{a}} = \underline{h}.$$

Using (3.7) in (3.8) we get

$$M'\Sigma^{-1}(\underline{\rho} - M\underline{\lambda}) = \underline{h}$$

whence

$$\underline{\lambda} = (M'\Sigma^{-1} M)^{-1}(M'\Sigma^{-1} \underline{\rho} - \underline{h})$$

and substituting this back in (3.7) we get finally

$$\underline{\tilde{a}} = \Sigma^{-1} \underline{\rho} - \Sigma^{-1} M(M'\Sigma^{-1} M)^{-1}(M'\Sigma^{-1} \underline{\rho} - \underline{h}).$$

The inversion of the matrices is justified since Σ is nonsingular and M is of full rank, due to the linear independence of the restraints (3.3). Whereas it is clear that $\underline{\tilde{a}}$ does not correspond to a maximum (i.e., F is unbounded as $|a_i| \longrightarrow \infty$) it is not obvious that $\underline{\tilde{a}}$ does not correspond to some other type of stationary point. We may argue geometrically to show this is not the case.

In the notation of Chapter 1, M_{n+h}^{t+h} is a vector space of n+h dimensions spanned by the elements $X_{t+h}, X_{t+h-1}, \ldots X_t, X_{t-1}, \ldots X_{t-n+1}$. The inner product is that defined by (1.5) and the norm by (1.6). Similarly, we have the space M_n^t. All elements of M_n^t whose coefficient vectors satisfy (3.3) form a subspace of M_n^t of n - m - 1 dimensions which we denote by V_n^t. Since V_n^t is a closed subspace of M_n^t and consequently of M_{n+h}^{t+h} there exists a unique projection of the element X_{t+h}

on V_n^t. But this in turn means that (3.4) has a unique minimum relative to $\underline{a}$ and subject to (3.5). Since our formal operations on $F(\underline{a},\underline{\lambda})$ give all stationary points of (3.4) subject to (3.5) this shows that (3.7) does correspond to a minimum.

We see that

$$\underline{\tilde{a}}'\underline{\rho} = \underline{\rho}'\underline{\tilde{a}} = \underline{\rho}'\Sigma^{-1}\underline{\rho} - \underline{\rho}'\Sigma^{-1}M(M'\Sigma^{-1}M)^{-1}(M'\Sigma^{-1}\underline{\rho} - \underline{h})$$

while

$$\underline{\tilde{a}}'\Sigma\underline{\tilde{a}} = [\underline{\rho}'\Sigma\underline{\rho} - 2\underline{\rho}'\Sigma^{-1}M(M'\Sigma^{-1}M)^{-1}(M'\Sigma^{-1}\underline{\rho} - \underline{h})$$

$$+ (M'\Sigma^{-1}\underline{\rho} - \underline{h})'(M'\Sigma^{-1}M)^{-1}(M'\Sigma^{-1}\underline{\rho} - \underline{h})]$$

and since the mean square error for any $\underline{a}$ satisfying (3.5) is given by (3.4) we finally get for the mean square error of prediction

$$\sigma^2\{1 - \underline{\rho}'\Sigma^{-1}\underline{\rho} + (M'\Sigma^{-1}\underline{\rho} - \underline{h})'(M'\Sigma^{-1}M)^{-1}(M'\Sigma^{-1}\underline{\rho} - \underline{h})\}.$$

It is noted that if we put M equal to the null matrix (no restrictions) then the above predictor is just the best linear predictor developed by Kolmogorov for the discrete case with zero means. It is further noted that nowhere in the derivation do we make use of the stationarity of the process so the above formulae are valid for arbitrary discrete processes with nonsingular covariance matrices (with the obvious notational changes, of course).

$\underline{P_1}$.

If one were to work through Hájek's theorem on the problem we have been discussing one could ultimately arrive at an explicit expression for the best predictor in the finite case. The expression turns out to be

$$(3.9) \qquad \underline{\tilde{a}} = \Sigma^{-1}\underline{\rho} - \Sigma^{-1}T(T'\Sigma^{-1}T)^{-1}T'\Sigma^{-1}\underline{\rho} + \Sigma^{-1}T(T'\Sigma^{-1}T)^{-1}\underline{t}$$

where Σ and $\underline{\rho}$ are as defined before and

$$\underset{n\times(m+1)}{T} = \begin{pmatrix} 1 & t & t^2 & \cdots & t^m \\ 1 & t-1 & (t-1)^2 & \cdots & (t-1)^m \\ \vdots & & & & \\ 1 & (t-n+1) & (t-n+1)^2 & \cdots & (t-n+1)^m \end{pmatrix}$$

$$\underset{(m+1)\times 1}{\underline{t}} = \begin{pmatrix} 1 \\ t+h \\ (t+h)^2 \\ \vdots \\ (t+h)^m \end{pmatrix} .$$

It would appear from this formula that the best predictor depends on t. We know, however, from previous results that we may replace T with the matrix M which does not involve t and get the same result. We now see that the reason this works is due to the existence of a nonsingular matrix B such that

$$(3.10) \qquad T = MB.$$

Putting this in Hájek's formula reduces his to ours. It will be recalled that if we were estimating the unknown parameters in the mean function by Markov estimation (weighted least squares) then these estimates would involve t. The estimates are in fact

$$(3.11) \qquad \underline{\hat{\alpha}} = (T'\Sigma^{-1}T)^{-1}T'\Sigma^{-1}\underline{X}$$

where $\underset{m+1\times 1}{\underline{\alpha}}$ is the vector of unknown regression parameters.

$\underline{P_2}$.

Suppose we have constructed the best linear predictor under the assumption that the degree of M_t is m when in fact it is m' < m. We must have, of course, that m < n.

Denote the vector of constants of the predictor so constructed as

$$(3.12) \qquad \underline{\tilde{a}}'(m) = (\tilde{a}_0(m),\ \tilde{a}_1(m),\ldots,\ \tilde{a}_{n-1}(m)).$$

It will be recalled that for unbiasedness we required that

$$\sum_{i=0}^{n-1} \tilde{a}_i(m) = 1$$

$$\sum_{i=0}^{n-1} i\ \tilde{a}_i(m) = -h$$

$$\sum_{i=0}^{n-1} i^m\ \tilde{a}_i(m) = (-h)^m$$

but this means that for any m' < m then $\underline{\tilde{a}}'(m)\underline{X}$ is an unbiased predictor of X_{t+h} if the degree of M_t is m'. There will, of course, be a difference in the mean square error.

$\underline{P_3 \text{ and } P_4}$.

The conditions under which we prove the assertions in P_3 and P_4 are given in terms of the spectral distribution function introduced in (2.8) as well as the correlation function of the process. These conditions are:

i) F(y) is absolutely continuous.

ii) the components of the vector $\underline{\rho}'\Sigma^{-1}T$ are bounded as $n \to \infty$.

We deal first with P_3. Suppose, first of all, that the mean function M_t is known. Then putting $Y_t = X_t - M_t$ the best linear predictor of Y_{t+h} may be constructed from the observed Y_t's and is well known to be

(3.14) $$Y_{t+h} = \underline{\rho}'\Sigma^{-1}\underline{Y}$$

where $\underline{Y}' = (X_t - M_t, X_{t-1} - M_{t-1}, \ldots, X_{t-n+1} - M_{t-n+1})$ so that the best linear predictor of X_{t+h} is given by

(3.15) $$\tilde{X}_{t+h} = \underline{\rho}'\Sigma^{-1}\underline{X} - \underline{\rho}'\Sigma^{-1}\underline{N} + M_{t+h}$$

where $\underline{N}' = (M_t, M_{t-1}, \ldots M_{t-n+1})$. It is expected that the quantity $-\underline{\rho}'\Sigma^{-1}\underline{N} + M_{t+h}$ is, in some way, estimated in the linear predictor when M_t is not known. That this is the case may be shown as follows:

Consider the model

(3.16) $$E(\underline{X}) = \underline{N}.$$

There exists a nonsingular matrix A such that $A\Sigma A' = I$ or $\Sigma = (A'A)^{-1}$ and putting

(3.17) $$\underline{Z} = A\underline{X}$$

we have

(3.18) $$\text{Cov.}(\underline{Z}) = I$$

so that

(3.19) $$E(\underline{Z}) = A\underline{N}$$

where the components of $\underline{Z}$ are uncorrelated. We know further that

(3.20) $$\underline{N} = M\underline{N}_d$$

where $\underline{N}_d' = (M_t, M_t^{(1)}, M_t^{(2)}, \ldots, M_t^{(m)})$, M is as before, and the $M_t^{(i)}$ are derivatives of M_t. Hence

$$E(\underline{Z}) = AM\underline{N}_d. \tag{3.21}$$

Now, estimating the $\underline{N}_d$ vector by ordinary least squares on the $\underline{Z}$ vector results in Markov estimates (or best linear estimates) in terms of the $\underline{X}$ vector. In the usual way we get the best estimate of $\underline{N}_d$ as

$$\begin{aligned} \hat{\underline{N}}_d &= (M'A'AM)^{-1}M'A'\underline{Z} \\ &= (M'\Sigma^{-1}M)^{-1}M'\Sigma^{-1}\underline{X} \end{aligned} \tag{3.22}$$

so the best estimate of $\underline{N}$ is

$$\hat{\underline{N}} = M(M'\Sigma^{-1}M)^{-1}M'\Sigma^{-1}\underline{X} \tag{3.23}$$

while that of M_{t+h} is

$$\hat{M}_{t+h} = \underline{h}'(M'\Sigma^{-1}M)^{-1}M'\Sigma^{-1}\underline{X} \tag{3.24}$$

and we see that $-\underline{\rho}'\Sigma^{-1}\underline{N} + M_{t+h}$ is best estimated by

$$-\underline{\rho}'\Sigma^{-1}\hat{\underline{N}} + \hat{M}_{t+h} = -(M'\Sigma^{-1}\underline{\rho} - \underline{h})'(M'\Sigma^{-1}M)^{-1}M'\Sigma^{-1}\underline{X} \tag{3.25}$$

which appears in the best linear predictor for unknown mean function. The variance of this quantity in (3.25) is

$$(M'\Sigma^{-1}\underline{\rho} - \underline{h})'(M'\Sigma^{-1}M)^{-1}(M'\Sigma^{-1}\underline{\rho} - \underline{h}). \tag{3.26}$$

Let $\tilde{X}_{t+h,n}$ be the best linear predictor based on $\underline{X}$ when the mean is known and $\hat{X}_{t+h,n}$ be the best linear predictor based on $\underline{X}$ when the mean is unknown. Then, since both these quantities have the same expectation, we have

$$(3.27) \qquad E\{\hat{X}_{t+h,n} - \tilde{X}_{t+h,n}\}^2 = \mathrm{Var.}\{\hat{X}_{t+h,n} - \tilde{X}_{t+h,n}\}.$$

Now we see that

$$\tilde{X}_{t+h,n} - \hat{X}_{t+h,n} = (M'\Sigma^{-1}\underline{\rho} - \underline{h})'(M'\Sigma^{-1}M)^{-1}M'\Sigma^{-1}(\underline{X} - \underline{N})$$

since $-\underline{\rho}'\Sigma^{-1}\underline{N} + M_{t+h}$ is unbiasedly estimated by

$$-(M'\Sigma^{-1}\underline{\rho} - \underline{h})'(M'\Sigma^{-1}M)^{-1}M'\Sigma^{-1}\underline{X}.$$

Then we get

$$(3.28) \quad \mathrm{Var.}\{\hat{X}_{t+h,n} - \tilde{X}_{t+h,n}\} = (M'\Sigma^{-1}\underline{\rho} - h)'(M'\Sigma^{-1}M)^{-1}(M'\Sigma^{-1}\underline{\rho} - h).$$

Thus $\hat{X}_{t+h,n} \longrightarrow \tilde{X}_{t+h,n}$ in the mean as $n \longrightarrow \infty$, if and only if

$$(3.29) \qquad (M'\Sigma^{-1}\underline{\rho} - \underline{h})'(M'\Sigma^{-1}M)^{-1}(M'\Sigma^{-1}\underline{\rho} - \underline{h}) \longrightarrow 0$$

as $n \longrightarrow \infty$. In view of (3.10), (3.29) may also be written

$$(3.30) \qquad (T'\Sigma^{-1}\underline{\rho} - \underline{t})'(T'\Sigma^{-1}T)^{-1}(T'\Sigma^{-1}\underline{\rho} - \underline{t}) \longrightarrow 0$$

as $n \longrightarrow \infty$.

Grenander and Rosenblatt [4] prove that the least squares estimates (not Markov) of the unknown regression parameters in a stationary time series with a polynomial mean function are mean consistent if the spectral distribution function satisfies i) above. Since the Markov estimates cannot have variances larger than the least squares estimates we have from (3.11)

$$\text{(3.31)} \qquad \text{Cov.}\ (\hat{\underline{\alpha}}) = (T'\Sigma^{-1}T)^{-1}$$

and the diagonal elements of $(T'\Sigma^{-1}T)^{-1}$ -- and hence all of them -- tend to zero as $n \longrightarrow \infty$. This implies that

$$\text{(3.32)} \qquad \underline{b}'(T'\Sigma^{-1}T)^{-1}\underline{b} \longrightarrow 0$$

as $n \longrightarrow \infty$ for any vector $\underline{b}$ whose components are bounded as $n \longrightarrow \infty$. Since the components of $\underline{t}$ are bounded and those of $T'\Sigma^{-1}\underline{\rho}$ are bounded by assumption ii) then (3.30) is true and hence $\hat{X}_{t+h,n} \longrightarrow \tilde{X}_{t+h,n}$ in the mean as $n \longrightarrow \infty$.

It is worth noting that conditions i) and ii) above hold for all stationary autoregressive schemes of finite order. This is because such processes have rational spectral densities and a fixed number of components (equal to the order of the scheme) of $\underline{\rho}'\Sigma^{-1}$ are nonzero and independent of n.

We now turn to property P_4. Suppose we have constructed the predictor assuming the degree of M_t to be m. Let $\Delta_m(m',n)$ be the mean square error of prediction based on n observations when M_t is of degree m' but has been assumed to be m. For $m' < m < n$

$$\text{(3.33)} \qquad \Delta_m(m',n) = E\{X_{t+h} - \sum_{i=0}^{n-1} \tilde{a}_i(m)X_{t-i}\}^2$$

$$= E\{(X_{t+h} - M_{t+h}) - \sum_{i=0}^{n-1} \tilde{a}_i(m)(X_{t-i} - M_{t-i})\}^2$$

due to the unbiasedness of the predictor, where the expectation is taken under the model where M_t is of degree m'. But (3.33) is seen to be just $\Delta_m(m,n)$, the mean square error of prediction when M_t is of degree m and we have guessed it correctly. Hence

$$(3.34)\quad \Delta_m(m',n) = \Delta_m(m,n)$$

$$= \sigma^2\{1 - \underline{\rho}'\ \Sigma^{-1}\underline{\rho} + (M'\Sigma^{-1}\underline{\rho} - \underline{h})'$$

$$(M'\Sigma^{-1}M)^{-1}(M'\Sigma^{-1}\underline{\rho} - \underline{h})\}.$$

If the degree of M_t is m' and we guess it correctly then the mean square error is

$$(3.35)\quad \Delta_{m'}(m',n) = \sigma^2\{1 - \underline{\rho}'\Sigma^{-1}\underline{\rho} + (M_1'\Sigma^{-1}\underline{\rho} - \underline{h}_1)'$$

$$(M_1'\Sigma^{-1}M_1)^{-1}(M_1'\Sigma^{-1}\underline{\rho} - \underline{h}_1)\}$$

where

$$\underset{(m'+1)\times n}{M_1'} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 2 & \cdots & (n-1) \\ 0 & 1 & 2^2 & \cdots & (n-1)^2 \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ 0 & 1 & 2^{m'} & \cdots & (n-1)^{m'} \end{pmatrix}$$

$$\underline{h}_1' = (1,-h,h^2,\ \ldots\ (-h)^{m'}).$$

It is seen that we may partition M and $\underline{h}$ as

$$M' = \begin{pmatrix} M_1' \\ \\ M_2' \end{pmatrix}$$

$$\underline{h} = \begin{pmatrix} \underline{h}_1 \\ \underline{h}_2 \end{pmatrix}$$

and since

$$M'\Sigma^{-1}\underline{\rho} - \underline{h} = \begin{pmatrix} M_1'\Sigma^{-1}\underline{\rho} - \underline{h}_1 \\ M_2'\Sigma^{-1}\underline{\rho} - \underline{h}_2 \end{pmatrix}$$

and

$$M'\Sigma^{-1}M = \begin{pmatrix} M_1'\Sigma^{-1}M_1 & M_1'\Sigma^{-1}M_2 \\ M_2'\Sigma^{-1}M_1 & M_2'\Sigma^{-1}M_2 \end{pmatrix}$$

it follows that

$$\Delta_m(m',n) - \Delta_{m'}(m',n) = \sigma^2\{(M'\Sigma^{-1}\underline{\rho} - \underline{h})'(M'\Sigma^{-1}M)^{-1}(M'\Sigma^{-1}\underline{\rho} - h)$$

$$- (M_1'\Sigma^{-1}\underline{\rho} - \underline{h}_1)(M_1'\Sigma^{-1}M_1)^{-1}(M_1'\Sigma^{-1}\underline{\rho} - \underline{h}_1)\}$$

$$\geq 0.$$

Therefore, guessing the degree of the polynomial too high increases the mean square error of prediction. We have noted, however, that when i) and ii) hold then (3.29) holds so that

$$\Delta_m(m',n) - \Delta_{m'}(m',n) \to 0$$

as $n \to \infty$. Hence, the increase in mean square error caused by guessing the degree of M_t too high diminishes as sample size increases.

We may express the difference in mean square error as a positive definite quadratic form with the help of the following lemma.

Lemma 3.1

If $\underset{m\times m}{S}$ is a positive definite matrix, $\underset{m\times 1}{\underline{x}}$ and $\underset{m\times 1}{\underline{y}}$ are column vectors and if they are partitioned as follows:

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

$$x = \begin{pmatrix} \underline{x}_1 \\ \underline{x}_2 \end{pmatrix}$$

$$\underline{y} = \begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \end{pmatrix}$$

then

$$\underline{x}'S^{-1}\underline{y} = \underline{x}_1'S_{11}^{-1}\ \underline{y}_1 + (\underline{x}_2 - S_{21}S_{11}^{-1}\underline{x}_1)'(S_{22}-S_{21}S_{11}^{-1}S_{12})^{-1}$$
$$(\underline{y}_2-S_{21}S_{11}^{-1}\underline{y}_1).$$

Proof:

Since $\underline{x}'\ S^{-1}\underline{y} = (A\underline{x})'\ (ASA')^{-1}\ (A\underline{y})$

for every nonsingular A, take

$$A = \begin{pmatrix} I & 0 \\ -S_{21}\ S_{11}^{-1} & I \end{pmatrix}$$

and the result follows.

If we use the lemma with

$$S = \begin{pmatrix} M_1'\Sigma^{-1}M_1 & M_1'\Sigma^{-1}M_2 \\ M_2'\Sigma^{-1}M_1 & M_2'\Sigma^{-1}M_2 \end{pmatrix}$$

$$\underline{x} = \underline{y} = \begin{pmatrix} M_1'\Sigma^{-1}\underline{\rho} - \underline{h}_1 \\ M_2'\Sigma^{-1}\underline{\rho} - \underline{h}_2 \end{pmatrix}$$

we see that

$$(3.36) \quad \Delta_m(m',n) - \Delta_{m'}(m',n) = \underline{y}_2' \, C^{-1} \, \underline{y}_2$$

where

$$\underline{y}_2 = M_2'\Sigma^{-1}\underline{\rho} - M_2'\Sigma^{-1}M_1(M_1'\Sigma^{-1}M_1)^{-1}(M_1'\Sigma^{-1}\underline{\rho} - \underline{h}_1) - \underline{h}_2$$

$$C = M_2'\Sigma^{-1}M_2 - (M_2'\Sigma^{-1}M_1)(M_1'\Sigma^{-1}M_1)^{-1}(M_1'\Sigma^{-1}M_2).$$

Examples of Best Predictors

Example 1

Suppose $m = 0$, i.e., constant mean, and the Y_t process is first order autoregressive. Then $\rho_{|i|} = \rho^{|i|}$ for $|\rho| < 1$ and we may find Σ^{-1} as follows:

Let

$$\underset{n\times n}{U} = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ 0 & 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}.$$

Then U^2 has 1's along the second upper diagonal and zeros elsewhere, U^3 has 1's along the third, etc. $U^n = 0$. Hence we may write

$$\underset{n\times n}{\Sigma} = \underset{n\times n}{I} + \rho(U + U') + \rho^2(U^2 + U'^2) + \ldots + \rho^{n-1}(U^{n-1} + U'^{n-1})$$

$$= I + \sum_{i=1}^{n-1} \rho^i (U^i + U'^i)$$

$$= (I - \rho U)^{-1} + \rho\, U'(I - \rho\, U')^{-1}$$

$$= (I - \rho\, U)^{-1}(I - \rho^2\, UU')(I - \rho\, U')^{-1}\ .$$

Since

$$(U\, U')^k = UU' \quad \text{(i.e., } UU' \text{ is idempotent)}$$

then

$$(I - \rho^2\, UU')^{-1} = I + UU'(\rho^2 + \rho^4 + \ldots)$$

$$= I + UU'(\rho^2/1 - \rho^2).$$

Therefore

$$\Sigma^{-1} = (I - \rho\, U')(I + \frac{\rho^2}{1-\rho^2}\, UU')(I - \rho\, U)$$

$$= I - \rho(U + U') + (\rho^2/1-\rho^2)(UU' + U'\, U)$$

$$- (\rho^3/1-\rho^2)(U'\, UU' + UU'\, U).$$

But

$$UU'U = U$$

and

$$U'\, UU' = U'$$

so

$$\Sigma^{-1} = I - (\rho/1-\rho^2)(U+U') + (\rho^2/1-\rho^2)(UU' + U'U),$$

a result which may now be verified directly.

Since

$$UU' = \begin{bmatrix} \underset{n-1\times n-1}{I} & 0 \\ 0 & 0 \end{bmatrix}$$

and

$$U'U = \begin{bmatrix} 0 & 0 \\ 0 & \underset{n-1\times n-1}{I} \end{bmatrix}$$

it follows that

$$\Sigma^{-1}\ \underline{\rho} = \begin{bmatrix} \rho^h \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

$$M'\Sigma^{-1}\ \underline{\rho} = \rho^h$$

$$\underset{\sim}{h} = 1$$

and

$$M'\Sigma^{-1}M = \{(1, 1, \ldots 1) - (\rho/1-\rho^2)(1, 2, 2, \ldots 2, 1)$$

$$+ (\rho^2/1-\rho^2)(1, 2, \ldots 2, 1)\}M$$

$$= \{(1, 1, \ldots 1) - (\rho/1+\rho)(1, 2, \ldots 2. 1)\} \begin{pmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix}$$

$$= (n - n\rho + 2\rho)/(1+\rho) \ .$$

Hence

$$\tilde{\underline{a}} = \begin{pmatrix} \rho^h \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix} + \frac{(1-\rho^h)(1+\rho)}{n-n\rho+2\rho} \begin{pmatrix} \frac{1}{1+\rho} \\ \frac{1-\rho}{1+\rho} \\ \cdot \\ \cdot \\ \cdot \\ \frac{1-\rho}{1+\rho} \\ \frac{1}{1+\rho} \end{pmatrix} \tag{3.37}$$

$$= \begin{pmatrix} \rho^h + \frac{(1-\rho^h)}{n-n\rho+2\rho} \\ \frac{(1-\rho^h)(1-\rho)}{n-n\rho+2\rho} \\ \cdot \\ \cdot \\ \cdot \\ \frac{(1-\rho^h)(1-\rho)}{n-n\rho+2\rho} \\ \frac{(1-\rho^h)}{n-n\rho+2\rho} \end{pmatrix} .$$

Thus our predictor is

$$\tilde{X}_{t+h,n} = \rho^h X_t + \frac{(1-\rho^h)(1-\rho)}{n-n\rho+2\rho} \sum_{i=1}^{n-2} X_{t-i} + \frac{(1-\rho^h)}{(n-n\rho+2\rho)} (X_t + X_{t-n+1})$$

with mean square error

(3.38) $$\sigma^2\{1 - \rho^{2h} + (1-\rho^h)^2(1+\rho)/(n-n\rho+2\rho)\}.$$

It is recalled that the best predictor in the case for known mean μ is

(3.39) $$\tilde{X}_{t+h} = \rho^h X_t + (1-\rho^h)\mu$$

with mean square error

(3.40) $$\sigma^2\{1 - \rho^{2h}\}$$

Example 2

Let $m = 0$ and $\rho_k = a\, p^k + b\, g^k$ for $k = 0, 1, 2, \ldots$

where

$$a = p(1-g^2)/(p-g)(1+pg)$$

$$b = g(1-p^2)/(g-p)(1+pg)$$

and both p and g are numerically less than 1. Then this correlation function is that of a stationary autoregressive scheme of order 2. Then

$$\begin{aligned}\Sigma &= I + (ap+bg)(U+U') + (ap^2+bg^2)(U^2+U'^2) \\ &\quad + \ldots, + (ap^{n-1}+bg^{n-1})(U^{n-1} + U'^{n-1}) \\ &= a\, I + a\left(\sum_{i=1}^{n-1} p^i U^i + \sum_{i=1}^{n-1} p^i U'^i\right) \\ &\quad + b\, I + b\left(\sum_{i=1}^{n-1} g^i U^i + \sum_{i=1}^{n-1} g^i U'^i\right)\end{aligned}$$

$$= a\{(I-pU)^{-1}(I+(p^2/1-p^2)UU')(I-p\,U')^{-1}\}$$

$$+ b\{(I-g\,U)^{-1}(I+(g^2/1-g^2)UU')(I-g\,U')^{-1}\}$$

$$= a\,J^{-1} + b\,K^{-1}, \text{ say.}$$

Now

$$a\,J^{-1} + b\,K^{-1} = a\,J^{-1}JJ^{-1}KK^{-1} + b\,J^{-1}JK^{-1}KK^{-1}$$

$$= J^{-1}(a\,K + b\,J)K^{-1}.$$

From the work with the first order scheme we have seen that

$$J = (I-pU')(I+(p^2/1-p^2)UU')(I-p\,U)$$

$$= I - (p/1-p^2)(U + U') + (p^2/1-p^2)(UU' + U'U)$$

and likewise,

$$K = I - (g/1-g^2)(U + U') + (g^2/1-g^2)(UU' + U'U).$$

Since

$$a\,K = \frac{p}{(p-g)(1+pg)}\{(1-g^2)I - g(U + U') + g^2(UU' + U'U)\}$$

and

$$b\,J = \frac{g}{(g-p)(1+pg)}\{(1-p^2)I - p(U + U') + p^2(UU' + U'U)\}$$

then

$$(a\,K + b\,J)^{-1} = \frac{(1+pg)}{(1-pg)}(I - pg\,R)$$

where

$$\underset{n\times n}{R} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

hence

$$\Sigma^{-1} = \frac{(1+pg)}{(1-pg)} K(I-pg\ R)J \ .$$

Some further calculations show that

$$M'\Sigma^{-1}M = \frac{(1+pg)((n-2)(1-g)(1-p) + 2(1-pg))}{(1-pg)(1+p)(1+g)}$$

and it is known that

$$\underline{\rho}'\Sigma^{-1} = \frac{1}{p-g}((p^{h+1}-g^{h+1}),\ pg(g^{h}-p^{h}),\ 0,\ 0,\ \ldots,\ 0)$$

since these are just the coefficients of the best predictor for known mean. Hence

$$\underline{\rho}\Sigma^{-1}M = \frac{1}{p-g}(p^{h+1} - g^{h+1} + pg^{h+1} - p^{h+1}g)$$

and finally we may write the coefficient vector as

$$(3.41)\quad \tilde{\underline{a}} = \frac{1}{p-g}\begin{pmatrix} p^{h+1}-g^{h+1} \\ pg^{h+1}-gp^{h+1} \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix}$$

$$+ \frac{(p-g-p^{h+1}+g^{h+1}-pg^{h+1}-gp^{h+1})}{(p-g)\{(n-2)(1-g)(1-p)+2(1-pg)\}} \times \begin{pmatrix} 1 \\ 1-p-g \\ (1-p)(1-g \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ (1-p)(1-g) \\ (1-p-g) \\ 1 \end{pmatrix}$$

The mean square error is now

$$(3.42) \quad \sigma^2\{1 - \frac{1}{(p-g)^2(1+pg)}((1-g^2)p^{2(h+1)}(1-pg) + (1-p^2)g^{2(h+1)}(1-pg) - 2p^{h+1}g^{h+1}(1-p^2)(1-g^2)) + \frac{(p-g-p^{h+1}+g^{h+1}-pg^{h+1}-gp^{h+1})^2(1-pg)(1+g)(1+p)}{(p-g)^2(1+pg)((n-2)(1-g)(1-p)+2(1-pg))}\}.$$

If g is set equal to zero then this case reduces to the first order result given earlier.

<u>Summary of Chapter 3</u>

In this section we have derived the best linear unbiased predictor of the value of a stochastic process at time t+h given its values at time t, t-1, ... t-n+1. We have found that the predictor is of the form of the best linear predictor when the mean is known with the weighted least squares estimates substituted for the unknown quantities. We have discovered that the predictor does not depend on t through the coefficients - only through the observations - that the predictor is unbiased even if we guess the degree of the polynomial mean function too high, and that the predictor does, under certain conditions, converge in the mean to the best linear predictor for known mean even though we have guessed the degree of M_t too high.

It is worth noting also that all the formulae thus far developed are valid for arbitrary discrete processes with nonsingular covariance matrix. The large sample results of P_3 and P_4, however, cannot be assured in such generality.

Chapter 4

PREDICTION FROM AN INFINITE PAST WHEN THE MEAN FUNCTION IS UNKNOWN

In this section we shall deal with the case where we have at our disposal the infinite past X_t, X_{t-1}, ... , and wish to investigate the existence of good predictors of the form

$$\sum_{i=0}^{\infty} a_i X_{t-i} . \tag{4.1}$$

When the process mean is known to be zero the assumption of an infinite past often leads to simpler formulae than for a finite past. Since the formulae are very complicated for a finite past when we do not know the mean we would hope the assumption of an infinite past simplifies our problem. Only real processes are studied.

It is well to remember that whether we know the mean function or not there always exists a best "linear" predictor in the sense that it is either a finite linear combination of the past values or a limit-in-mean of a sequence of such finite linear combinations. The case of known mean was established by Kolmogorov and that of unknown mean by Hájek. It should further be recalled, however, that this does not imply the predictor will be of the form we are considering. It is often the case when the mean is known, as we found in Chapter 2, but cannot be assured even then. When the mean is known, for instance, all finite order autoregressive schemes have best "linear" predictors depending only on a finite number of past observations.

In the case of unknown means we will first investigate the existence of a predictor, $\hat{X}_{t+h}$, subject to:

$$C_1^*.\ \hat{X}_{t+h} = \sum_{i=0}^{\infty} a_i X_{t-i}$$

$$C_2^*.\ E(\hat{X}_{t+h}) = E(X_{t+h}),\ t \varepsilon T$$

(4.2)

$$C_3^*.\ E\{\hat{X}_{t+h} - X_{t+h}\}^2 \text{ to be minimized}$$

$$C_4^*.\ E\{X_{t+h}\} = M_{t+h}$$

where M_t is a polynomial of degree m. These conditions are recognized as those of (1.18) with $n = \infty$. We shall mean by $\sum_{i=0}^{\infty} a_i X_{t-i}$ the infinite sum which is the limit-in-mean of the sequence $\{\sum_{i=0}^{n} a_i X_{t-i}\}$, n = 0, 1, 2, ... More specifically, we require

$$\lim_{n\to\infty} E\{\sum_{i=0}^{n} a_i X_{t-i} - \sum_{i=0}^{\infty} a_i X_{t-i}\}^2 = 0. \tag{4.3}$$

Let $\tilde{X}_{t+h}$ be the best "linear" predictor (in the classical sense) of X_{t+h} based on an infinite past when the mean is known. Let A_n be n dimensional vector space with elements $\underline{a}_n = (a_0, a_1, \ldots, a_{n-1})$, where the a_i are real, and let A_n^0 be that subspace of A_n for which

$$\sum_{i=0}^{n-1} a_i = 1. \tag{4.4}$$

Denote the extensions to infinite dimensional vector spaces as A_∞ and A_∞^0 respectively with the additional requirement that A_∞ contain only those sequences such that (4.3) is true. We first prove the following lemma.

Lemma 4.1

Suppose the stationary time series X_t obeys the model in (1.1) with $M_t = \mu$, $|\mu| < \infty$. Suppose further that for the sequence $a_0, a_1, a_2, \ldots$, (4.3) is true. Then we may conclude

i). $E\{\sum_{i=0}^{\infty} a_i X_{t-i}\} = \sum_{i=0}^{\infty} a_i E(X_{t-i}) = \mu \sum_{i=0}^{\infty} a_i$

ii). $\mathrm{Var}\{\sum_{i=0}^{\infty} a_i X_{t-i}\} = \lim_{n\to\infty} \sigma^2 \sum_{i=0}^{n} \sum_{j=0}^{n} a_i a_j \rho_{|i-j|}$.

Proof:

We note at the beginning that $\sum_{i=0}^{\infty} a_i X_{t-i}$, being a limit-in-mean, has a finite variance and hence the quantities in i) and ii) are finite.

From (4.3) we notice that in view of

$$E\{\sum_{i=0}^{\infty} a_i X_{t-i} - \sum_{i=0}^{n-1} a_i X_{t-i}\}^2$$

$$= E\{\sum_{i=0}^{\infty} a_i X_{t-i} - E(\sum_{i=0}^{\infty} a_i X_{t-i}) - \sum_{i=0}^{n-1} a_i X_{t-i} + \mu \sum_{i=0}^{n-1} a_i\}^2$$

$$+ \{E(\sum_{i=0}^{\infty} a_i X_{t-i}) - \mu \sum_{i=0}^{n-1} a_i\}^2$$

and since the left hand side tends to zero as $n\to\infty$ both the terms on the right must tend to zero. In particular,

$$|E(\sum_{i=0}^{\infty} a_i X_{t-i}) - \mu \sum_{i=0}^{n-1} a_i|^2 \to 0$$

as $n\to\infty$, or

$$E(\sum_{i=0}^{\infty} a_i X_{t-i}) = \lim_{n\to\infty} \mu \sum_{i=0}^{n-1} a_i = \mu \sum_{i=0}^{\infty} a_i \quad .$$

Clearly, (4.3) implies that

$$(4.5) \qquad E\{\sum_{i=0}^{\infty} a_i(X_{t-i}-\mu) - \sum_{i=0}^{n-1} a_i(X_{t-i}-\mu)\}^2 \to 0$$

as $n \to \infty$. But the left hand side of (4.5) is recognized as the square of the norm defined in (1.6) for the Hilbert space $\mathcal{H}$ (the $(X_{t-i}-\mu)$'s here correspond to the X_{t-i}'s there).

Since for norms we have the relation

$$(4.6) \qquad \Big| ||X|| - ||Y|| \Big| \leq ||X-Y||$$

then

$$|(E\{\sum_{i=0}^{\infty} a_i(X_{t-i}-\mu)\}^2)^{\frac{1}{2}} - (E\{\sum_{i=0}^{n-1} a_i(X_{t-i}-\mu)\}^2)^{\frac{1}{2}}|$$

$$\leq (E\{\sum_{i=0}^{\infty} a_i(X_{t-i}\ \mu) - \sum_{i=0}^{n-1} a_i(X_{t-i}-\mu)\}^2)^{\frac{1}{2}}.$$

Since the right hand side tends to zero as $n \to \infty$ then

$$E\{\sum_{i=0}^{\infty} a_i(X_{t-i}-\mu)\}^2 = \lim_{n\to\infty} E\{\sum_{i=0}^{n-1} a_i(X_{t-i}-\mu)\}^2$$

or

$$\mathrm{Var}\{\sum_{i=0}^{\infty} a_i X_{t-i}\} = \lim_{n\to\infty} \sigma^2 \sum_{i=0}^{n} \sum_{j=0}^{n} a_i a_j \rho_{|i-j|}.$$

<u>Theorem 4.1</u>

Suppose for a time series as in Lemma 4.1 we have

$$(4.7) \qquad (\underline{j}' \Sigma^{-1} \underline{\rho} - 1)^2/(\underline{j}' \Sigma^{-1} \underline{j}) \to 0$$

as $n \to \infty$, where $\underline{j}' = (1, 1, \ldots, 1)$, and if

$$(4.8) \qquad \tilde{X}_{t+h} = \sum_{i=0}^{\infty} b_i X_{t-i} + \mu(1 - \sum_{i=0}^{\infty} b_i)$$

then

$$\sum_{i=0}^{\infty} b_i \neq 1. \tag{4.9}$$

Then there does not exist a predictor of X_{t+h}, $\hat{X}_{t+h}$, which satisfies conditions (4.2) above when the process mean is an unknown constant μ.

Proof: Suppose an $\hat{X}_{t+h}$ satisfying (4.2) does exist. (4.7) is recognized as (3.29) for the special case where $M_t = \mu$ and so from Theorem 3.1, (P_3) we have

$$\begin{aligned}\lim_{n\to\infty} E\{X_{t+h} - \hat{X}_{t+h,n}\}^2 &= E\{X_{t+h} - \hat{X}_{t+h}\}^2 \\ &= \lim_{n\to\infty} \sigma^2(1 - \underline{\rho}'\ \Sigma^{-1}\ \underline{\rho}) \\ &= \sigma_0^2, \text{ say.}\end{aligned}$$

where $\hat{X}_{t+h,n}$, $\underline{\rho}'$, Σ^{-1}, are as defined before.

Further,

$$E\{X_{t+h} - \tilde{X}_{t+h}\}^2 = \sigma_0^2 .$$

If $\tilde{X}_{t+h}$ is not of the form in (4.8) above we reach an immediate contradiction due to

$$\begin{aligned}E\{X_{t+h} - \hat{X}_{t+h}\}^2 &= \inf_{\underline{a}^\infty \varepsilon A_\infty^0} E\{X_{t+h} - \sum_{i=0}^{\infty} a_i\ X_{t-i}\}^2 \\ &= \inf_{\underline{a}^\infty \varepsilon A_\infty^0} E\{(X_{t+h}-\mu) - \sum_{i=0}^{\infty} a_i(X_{t-i}-\mu)\}^2 \\ &= E\{(X_{t+h}-\mu) - \sum_{i=0}^{\infty} a_i^*(X_{t-i}-\mu)\}^2 \\ &= \sigma_0^2\end{aligned}$$

for some $\underline{a}_\infty^* \ \varepsilon\ A_\infty^0$. This is because

$$\sum_{i=0}^{\infty} a_i = 1$$

for $\underline{a}_\infty \in A_\infty^0$. We also have used i) for lemma 4.1 in these calculations. But

$$E\{(X_{t+h}-\mu) - \sum_{i=0}^{\infty} a_i^*(X_{t-i}-\mu)\}^2 = \sigma_0^2$$

means that

$$\tilde{X}_{t+h} = \sum_{i=0}^{\infty} a_i^* X_{t-i} + \mu(1 - \sum_{i=0}^{\infty} a_i^*)$$

which is a contradiction.

If $\tilde{X}_{t+h}$ is of the form in (4.8) then we have

$$\tilde{X}_{t+h} = \sum_{i=0}^{\infty} b_i X_{t-i} + \mu(1 - \sum_{i=0}^{\infty} b_i)$$

for some sequence of b_i's and also

$$\tilde{X}'_{t+h} = \sum_{i=0}^{\infty} a_i^* X_{t-i} + \mu(1 - \sum_{i=0}^{\infty} a_i^*)$$

so they must be equivalent in the sense that

$$E\{\tilde{X}'_{t+h} - \tilde{X}_{t+h}\}^2 = 0$$

which means that

$$E\{\sum_{i=0}^{\infty}(b_i - a_i^*)(X_{t-i}-\mu)\}^2 = 0$$

and from ii) of lemma 4.1 we have

$$\lim_{n\to\infty} \sigma^2 \sum_{i=0}^{n} \sum_{j=0}^{n}(a_i^* - b_i)(a_j^* - b_j)\, \rho_{|i-j|} = 0.$$

This means, however, that

$$a_i^* = b_i \qquad\qquad i = 0, 1, 2, \ldots,$$

and since $\sum_{i=0}^{\infty} b_i \neq 1$ this is a contradiction.

It will be recalled that in the first order autoregressive scheme we have for $M_t = \mu$

$$\tilde{X}_{t+h} = \rho^h X_t + (1 - \rho^h)\mu$$

and $(\underline{j}' \Sigma^{-1} \underline{\rho} - 1)^2/(\underline{j}' \Sigma^{-1} \underline{j}) = (1 - \rho^h)^2(1 + \rho)/n - n\rho + 2\rho$.

Since

$$(\rho^h, 0, 0, \ldots) \notin A_\infty^0$$

then the hypothesis is fulfilled and in this case we can never get a best predictor which is unbiased and linear in the observations when we do not know the mean. A natural inquiry is whether there are cases where the coefficient vector $\underline{b}_\infty$ of the best "linear" predictor, $\tilde{X}_{t+h}$, does lie in A_∞^0. If such cases exist we automatically get a solution to our problem.

It is seen then that we must abandon some of the requirements set out above for our predictor and since in practice it is generally sufficient to have a small mean square error, we shall look for predictors with this property. Before proceeding, however, we point out that Hâjek's solution to the problem of finding a best predictor based on an infinite past is far from a practical solution. First of all, the solution is unfortunately a limit-in-mean of a sequence of linear functions of the observations and secondly, the solution generally involves the unknown parameters in the model. This is seen by noting that if (3.29) holds then we must have

$$\lim_{n\to\infty} E\{\hat{X}_{t+h,n} - X_{t+h}\}^2 = \sigma_0^2$$

$$= E\{\tilde{X}_{t+h} - X_{t+h}\}^2.$$

This follows from

$$\lim_{n\to\infty} E\{\hat{X}_{t+h,n} - \tilde{X}_{t+h}\}^2 = 0$$

and hence $\tilde{X}_{t+h}$ is the limit-in-mean of a sequence of linear functions of the past observations which achieves the minimum mean square error, σ_0^2. Then $\tilde{X}_{t+h}$ is a solution to the problem in the Hájek formulation and is the only solution. $\tilde{X}_{t+h}$, however, involves the unknown mean function as we have seen for the single autoregressive case.

There are in general several ways of constructing predictors which are arbitrarily close to the best when we have an infinite past. One way is to take an arbitrarily large portion of the observations and construct the best predictor for that number. This, however, defeats the very purpose of assuming an infinite past. There are cases where we can use all the observations and construct a relatively simple predictor which has mean square error arbitrarily close to the minimum and at the same time enjoys other nice properties. D. R. Cox [3] discusses a very simple case in connection with his moving average predictor. In deriving an "almost" optimum predictor as Cox does in a very special case, the basic idea is to take the optimum predictor for known mean and replace the mean with an estimate of moving average type. Cox discusses the first order autoregressive case in some detail and we reiterate some of his ideas before generalizing his results.

Suppose we consider the special case of our model where

(4.10) a) $M_t = \mu$, $t = 0, \pm 1, \pm 2, \ldots$

b) $\sigma^2 \rho_{|k|} = \sigma^2 \rho^{|k|}$, $k = 0, \pm 1, \pm 2, \ldots$

and $|\rho| < 1$. We wish to predict X_{t+h} by a linear function of the past of the form

$$(4.11) \qquad \hat{X}_{t+h} = (1 - \lambda) \sum_{i=0}^{\infty} \lambda^i X_{t-i} , \qquad |\lambda| < 1.$$

This type of predictor is called an exponentially weighted moving average (ewma) and is easily shown to satisfy conditions C_1^*, C_2^*, C_4^* above. Denote the mean square error of such a predictor as $\Delta_c(h, \lambda)$ when we are predicting h steps ahead. The principle advantages of such a predictor are:

(4.12) i) The mean μ need not be known

ii) If long term drifts in the mean occur, the predictor will follow such a trend.

Routine calculation reveals that

$$(4.13) \quad \Delta_c(h, \lambda) = 2\sigma^2\{\frac{1}{1+\lambda} + \frac{(1-\lambda)}{1+\lambda} \sum_{r=1}^{\infty} \lambda^r \rho^r - (1-\lambda) \sum_{r=0}^{\infty} \lambda^r \rho^{h+r}\}$$

$$= \frac{2\sigma^2}{1+\lambda}\{1 + \frac{(1-\lambda)(\lambda\rho-\rho^h-\lambda\rho^h)}{(1-\lambda\rho)}\}$$

and for the special case where h = 1 we have

$$(4.14) \qquad \Delta_c(1, \lambda) = \frac{2\sigma^2(1-\rho)}{(1-\lambda\rho)(1+\lambda)} .$$

Hence $\Delta_c(1, \lambda)$ will be minimized relative to λ, remembering that $|\lambda| < 1$, by

$$(4.15) \qquad \lambda_0 = \frac{(1-\rho)}{2\rho} \qquad \text{if } \rho > 1/3$$

$$= 1 \qquad \text{if } \rho \leq 1/3$$

where the case $\lambda = 1$ may be defined as meaning we take a simple unweighted mean of the observations sufficient to determine μ

without appreciable error. The corresponding mean square errors are:

$$(4.16) \qquad \Delta_c(1, \lambda_0) = \frac{8\sigma^2\rho(1-\rho)}{(1+\rho)^2} \qquad \text{if } \rho > 1/3$$

$$= \sigma^2 \qquad \text{if } \rho \leq 1/3.$$

So we observe this predictor is of little use in the simple autoregressive case when $\rho \leq 1/3$. It can be shown that this critical value of ρ in fact increases as we predict farther into the future. A few of these critical values are:

(4.17)

h	1	2	4
Critical ρ	0.333	0.516	0.821

Various other interesting features of this predictor for the simplest autoregressive case are presented in Cox's paper [3] and a comparison is made with the best linear predictor when the mean is known.

Our aim here is to first of all examine the finite sample-size version of Cox's predictor for the simple autoregressive case to see how important sample size is, to show that the Cox predictor may be useful in more general settings, and to show how a Cox-type predictor may be used in constructing almost optimum linear predictors.

<u>Lemma 4.2</u>

If

$$\hat{X}_{t+h,n} = \frac{(1-\lambda)}{(1-\lambda^n)} \sum_{i=0}^{n-1} \lambda^i X_{t-i}$$

then the mean space error, $\Delta_c(h, \lambda, n)$, is

$$\Delta_c(h, \lambda, n) = \frac{\sigma^2(1-\lambda)^2}{(1-\lambda^n)^2}\left\{\frac{(1-\lambda^n)^2}{(1-\lambda)^2} - \frac{2(1-\lambda^n)}{(1-\lambda)}\sum_{i=0}^{n-1} \rho_{h+i}\, \lambda^i + \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} \lambda^{i+j}\, \rho_{|i-j|}\right\}$$

and if $\rho_{|k|} = \rho^{|k|}$ we get

$$\Delta_c(h, \lambda, n) = \frac{2\sigma^2(1-\lambda)^2}{(1-\lambda^n)^2}\left\{\frac{(1-\lambda^n)(1-\lambda^{n+1})}{(1-\lambda)^2(1+\lambda)} + \frac{\lambda\rho}{1-\lambda\rho}\left(\frac{1-\lambda^{2(n-1)}}{1-\lambda^2} - \frac{(\lambda\rho)^{n-1}(1-(\frac{\lambda}{\rho})^{n-1})}{1-\frac{\lambda}{\rho}}\right) - \frac{\rho^h(1-\lambda^n)(1-(\lambda\rho)^n)}{(1-\lambda)(1-\lambda\rho)}\right\}$$

if $\lambda \neq \rho$ and

$$\Delta_c(h, \rho, n) = \lim_{\lambda\to\rho} \Delta_c(h, \lambda, n).$$

Proof:

The calculations are routine.

This expression, $\Delta_c(n, \lambda, n)$ is rather complicated even if $h = 1$ and minimization relative to λ is formidable. We may, however, make some numerical comparisons, keeping in mind that the Cox predictor makes no claim of goodness if $\rho \leq 1/3$.

The mean square errors of the following four predictors are compared for sample sizes 2, 4, 6, and 10, when the process is the simple autoregressive one discussed above.

Let

a) $\Delta_c(h, \lambda_0)$ = the mean square error of Cox's predictor using the optimum λ, λ_0 say.

b) $\Delta_c(h, \lambda_0, n)$ = the mean square error of the finite sample size version of Cox's predictor, using λ_0 as determined from $\Delta_c(h, \lambda)$.

c) $\Delta(h, \text{opt.}, n)$ = the mean square error of the best linear predictor for sample size n and unknown mean μ (Chapter 2, example 1).

d) $\Delta(h, \text{opt.})$ = the mean square error of the best linear predictor for the case of known mean μ.

The tabulations (pages 80 and 81) are made only for the case $h = 1$ and the tabled quantity is the mean square error divided by σ^2, the variance of the process.

The principle points illustrated by these tables are, first of all, the ewma performs quite well if the correlation between adjacent observations is high (greater than .70) and secondly, the ewma is essentially based on the last ten observations even though it is a function of the entire past. The last feature assures us, of course, that any shift in the mean occurring in advance of the past ten observations may be ignored since the predictor is insensitive to such shifts.

We now show that Cox's predictor may be used in other settings with some assurance that it will never be worse than using the mean of the process (unweighted average of the past) as a predictor.

Define for $\lambda \in (0,1)$ and h a fixed positive integer, the following:

$$f(\lambda) = 1/1+\lambda$$

$$g(\lambda) = \{\sum_{i=0}^{\infty} \lambda^{i}(\rho_i-\rho_{h+i}) - \sum_{i=0}^{\infty} \lambda^{i+2}(\rho_{i+1}-\rho_{h+i})\}$$

n=2

ρ	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
a	.998	.998	.942	.888	.824	.750	.668	.582	.490	.396	.298	.198	.100
b	1.194	1.112	1.028	.944	.858	.770	.680	.586	.492	.396	.298	.200	.100
c	1.163	1.092	.944	.938	.826	.768	.679	.586	.510	.396	.298	.200	.100
d	.878	.840	.798	.750	.698	.640	.578	.510	.438	.360	.278	.190	.098

n=4

ρ	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
a	.998	.998	.942	.888	.824	.750	.668	.582	.490	.395	.298	.199	.100
b	1.134	1.056	.982	.910	.834	.754	.670	.582	.490	.396	.298	.200	.100
c	1.051	.998	.905	.875	.806	.731	.653	.567	.482	.390	.296	.199	.100
d	.878	.840	.798	.750	.698	.640	.578	.510	.438	.360	.278	.190	.098

n=6

ρ	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
a	.998	.998	.942	.888	.824	.750	.668	.582	.490	.395	.298	.199	.100
b	1.098	1.024	.960	.896	.826	.750	.668	.582	.490	.396	.298	.199	.100
c	1.002	.955	.882	.844	.781	.711	.637	.558	.474	.386	.294	.198	.100
d	.878	.840	.798	.750	.698	.640	.578	.510	.438	.360	.278	.190	.098

n=10

ρ	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
a	.998	.998	.942	.888	.824	.750	.668	.582	.490	.395	.298	.199	.100
b	1.056	.994	.944	.890	.824	.750	.668	.582	.490	.396	.298	.199	.100
c	.957	.914	.857	.812	.754	.689	.620	.545	.465	.380	.291	.197	.100
d	.878	.840	.798	.750	.698	.640	.578	.510	.438	.360	.278	.190	.098

where ρ_i is the autocorrelation function of a stationary time series. Then we may state the following result.

<u>Theorem 4.2</u>

Suppose $g''(\lambda) > 0$ on $(0,1)$, $\sum_{i=0}^{\infty} |\rho_i| < \infty$, and $\rho_h < 1$, then the predictor

$$\hat{X}_{t+h} = (1-\lambda) \sum_{i=0}^{\infty} \lambda^i X_{t-i} , \qquad \lambda \in (0,1)$$

has the following properties:

(4.18)

p_1) Its mean square error is given by

$$\Delta(h, \lambda) = 2\sigma^2 f(\lambda)\, g(\lambda).$$

p_2) $\Delta(h, \lambda)$ has a unique minimum at a point λ_0 in $(0,1)$ if either

i) $\rho_1 + \rho_h - \rho_{h+1} < 1$ <u>and</u> $\left(\sum_{i=h}^{\infty} \rho_i - \sum_{i=1}^{h-1} \rho_i\right) > \frac{1}{2}$

or

ii) $\rho_1 + \rho_h - \rho_{h+1} > 1$ <u>and</u> $\left(\sum_{i=h}^{\infty} \rho_i - \sum_{i=1}^{h-1} \rho_i\right) < \frac{1}{2}$

where $\sum_{i=1}^{0} \rho_i \equiv 0$.

p_3) The minimum mean square error is

$$\Delta(h, \lambda_0) = 2\sigma^2 g'(\lambda_0).$$

Proof:

p_1). The mean square error of prediction for any positive integral h is

$$E\{X_{t+h} - \hat{X}_{t+h}\}^2 = (1-\lambda)^2 E\{\sum_{i=0}^{\infty} \lambda^i (X_{t+h}-X_{t-i})\}^2$$

$$= (1-\lambda)^2 \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \lambda^{i+j} E\{(X_{t+h}-X_{t-i})(X_{t+h}-X_{t-j})\}$$

$$= \sigma^2(1-\lambda)^2 \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \lambda^{i+j} (1-\rho_{h+i}-\rho_{h+j}+\rho_{|i-j|})$$

$$= \sigma^2(1-\lambda)^2 \left\{\frac{1}{(1-\lambda)^2} - \frac{2}{(1-\lambda)} \sum_{i=0}^{\infty} \lambda^i \rho_{h+i} + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \lambda^{i+j} \rho_{|i-j|}\right\}.$$

But

$$\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \lambda^{i+j} \rho_{|i-j|} = \sum_{i=0}^{\infty} \lambda^{2i} + 2 \sum_{i=0}^{\infty} \sum_{k=1}^{\infty} \lambda^{2i+k} \rho_k$$

$$= \frac{1}{1-\lambda^2} + \frac{2}{1-\lambda^2} \sum_{k=1}^{\infty} \lambda^k \rho_k$$

so that

$$\Delta(h, \lambda) = \sigma^2(1-\lambda)^2 \left\{\frac{1}{(1-\lambda)^2} - \frac{2}{(1-\lambda)} \sum_{i=0}^{\infty} \lambda^i \rho_{h+i} + \frac{1}{1-\lambda^2} (1 + 2 \sum_{k=1}^{\infty} \lambda^k \rho_k)\right\}$$

$$= \frac{2\sigma^2}{1+\lambda} \left\{1-(1-\lambda^2) \sum_{i=0}^{\infty} \lambda^i \rho_{h+i} + (1-\lambda) \sum_{r=1}^{\infty} \lambda^r \rho_r\right\}$$

$$= \frac{2\sigma^2}{1+\lambda} \left\{\sum_{i=0}^{\infty} \lambda^i (\rho_i-\rho_{h+i}) - \sum_{i=0}^{\infty} \lambda^{i+2} (\rho_{i+1}-\rho_{h+i})\right\}$$

$$= 2\sigma^2 f(\lambda) g(\lambda).$$

p_2). Put $\qquad F(\lambda) = f(\lambda)\, g(\lambda)$

and
$$\begin{aligned} F_1(\lambda) &= \log F(\lambda) \\ &= \log f(\lambda) + \log g(\lambda) \\ &= f_1(\lambda) + g_1(\lambda), \text{ say.} \end{aligned}$$

$\Delta(h, \lambda)$ has a unique minimum on (0,1) if and only if $F_1(\lambda)$ does and $F_1(\lambda)$ has a unique minimum on (0,1) if and only if there exists exactly one point λ_0 in (0,1) such that

$$F_1'(\lambda_0) = 0$$

$$F_1''(\lambda_0) > 0.$$

Since $\qquad F_1'(\lambda_0) = 0$

means that $\qquad f_1'(\lambda_0) = -g'(\lambda_0)$

or

$$\begin{aligned} \frac{1}{1+\lambda_0} &= g_1'(\lambda_0) \\ &= \frac{g'(\lambda_0)}{g(\lambda_0)} \end{aligned}$$

then we have

$$\begin{aligned} F_1''(\lambda_0) &= \frac{1}{(1+\lambda_0)^2} + g_1''(\lambda_0) \\ &= \left[\frac{g'(\lambda_0)}{g(\lambda_0)}\right]^2 + \frac{g''(\lambda_0)}{g(\lambda_0)} - \left[\frac{g'(\lambda_0)}{g(\lambda_0)}\right]^2 \\ &= \frac{g''(\lambda_0)}{g(\lambda_0)} \\ &> 0. \end{aligned}$$

This means any solution of

(4.19) $\qquad f_1'(\lambda) = -g_1'(\lambda)$

must correspond to a minimum and hence there can be at most one solution. The fact that there is a solution follows from

$$f_1'(0) = -1$$

$$f_1'(1) = -\tfrac{1}{2}$$

$$g_1'(0) = (\rho_1 - \rho_{h+1})/(1-\rho_h)$$

$$g_1'(1) = (\sum_{i=h}^{\infty} \rho_i - \sum_{i=1}^{h-1} \rho_i)$$

which means (using 4.18) either

$$f_1'(0) < -g_1'(0)$$

$$f_1'(1) > -g_1'(1)$$

or

$$f_1'(0) > -g_1'(0)$$

$$f_1'(1) < -g_1'(1)$$

and hence there must be a solution for (4.19).

P_3.

If a λ_0 exists which satisfies (4.19) then

$$f_1'(\lambda_0) = -g_1'(\lambda_0)$$

$$-\frac{1}{1+\lambda_0} = -\frac{g'(\lambda_0)}{g(\lambda_0)} \; .$$

Hence

$$\frac{g(\lambda_0)}{1+\lambda_0} = g(\lambda_0)\ f(\lambda_0)$$

$$= g'(\lambda_0)$$

and

$$\Delta(h, \lambda_0) = 2\sigma^2 g'(\lambda_0) .$$

Theorem 4.2a

If $h = 1$, $\rho_i \geq \rho_{i+1}$ for all i, $\sum_{i=0}^{\infty} i\, \rho_i$ is convergent, and $\rho_1 < 1$ then the conditions of Theorem 4.2 are satisfied.

Conditions (4.18) reduce to

(4.20) i) $\rho_1 < (1+\rho_2)/2$ and $\sum_{i=1}^{\infty} \rho_i > \frac{1}{2}$.

When (4.20) holds we may bound $\Delta(1,\lambda)$ below and above respectively in a nontrivial way by:

$$B_\ell(\lambda) = 2\sigma^2(a_\ell\, \lambda^2 + b_\ell\, \lambda + c_\ell) \tag{4.21}$$

$$B_u(\lambda) = 2\sigma^2 \left(\frac{1}{2(1-\rho_1)}\right)^{\lambda} (1-\rho_1)$$

where

$$a_\ell = -(\rho_1 - \rho_2)/2$$

$$b_\ell = \sqrt{2}\, \rho_1 - (\sqrt{2} - \tfrac{1}{2})\, \rho_2 - \tfrac{1}{2}$$

$$c_\ell = (\sqrt{2} - \tfrac{1}{2})(1 - \rho_1).$$

In addition we have

$$2\sigma^2(\rho_1 - \rho_2) \leq \Delta(1, \lambda_0) \leq \min\{2\sigma^2(1 - \rho_1), \sigma^2\}. \tag{4.22}$$

Proof:

For $h = 1$,

$$g(\lambda) = \sum_{i=0}^{\infty} \lambda^i (\rho_i - \rho_{i+1})$$

and since $(\rho_i - \rho_{i+1}) \geq 0$ for all i, $g(\lambda)$ is clearly convex on (0,1).

Also it is clear that the ρ_i are positive so that the convergence of $\sum_{i=0}^{\infty} i\, \rho_i$ implies that

$$\sum_{i=0}^{\infty} \rho_i = \sum_{i=0}^{\infty} |\rho_i|$$

$$< \infty$$

and the conditions of Theorem 4.2 are satisfied.

The i) part of (4.18) obviously reduces to i) of (4.20) if $h = 1$ since $\sum_{i=1}^{0} \rho_i \equiv 0$. On the other hand, ii) of (4.18) reduces to

$$\rho_1 > (1 + \rho_2)/2 \text{ and } \sum_{i=1}^{\infty} \rho_i < \tfrac{1}{2}.$$

But the first of these statements means

$$\rho_1 - \rho_2/2 > \tfrac{1}{2}$$

and since the ρ_i are positive this is incompatible with the second statement.

Then ii) of (4.18) cannot hold for the hypothesis of theorem 4.2a.

Now

$$F_1(0) = \log f(0) + \log g(0)$$

$$= \log (1 - \rho_1)$$

$$F_1(1) = \log \tfrac{1}{2}.$$

Since $F_1(\lambda)$ is decreasing on $(0, \lambda_0)$ and increasing on $(\lambda_0, 1)$ when (4.20) holds it is bounded above on $(0, 1)$ by the line

$$\lambda[F_1(1) - F_1(0)] + F_1(0) = \lambda(\log \tfrac{1}{2} - \log(1 - \rho_1)) + \log(1 - \rho_1)$$

$$= \log \{(\frac{1}{2(1-\rho_1)})^{\lambda} (1-\rho_1)\}$$

so that

$$F(\lambda) \leq (\frac{1}{2(1-\rho_1)})^{\lambda} (1-\rho_1)$$

$$= B_u(\lambda), \text{ say.}$$

Further,

$$B_u''(\lambda) = (1 - \rho_1)(\log \frac{1}{2(1-\rho_1)})^2 (\frac{1}{2(1-\rho_1)})^{\lambda}$$

$$\geq 0$$

and, of course,

$$B_u(0) = F(0)$$

$$B_u(1) = F(1)$$

so $B_u(\lambda)$ is a convex upper bound of $F(\lambda)$ which equals $F(\lambda)$ at the endpoints.

$F(\lambda)$ may be bounded below in a nontrivial way by noting that both $f(\lambda)$ and $g(\lambda)$ are convex on $(0,1)$ and hence may be bounded below by lines.

For the line bounding $f(\lambda)$ we choose that line parallel to the segment joining the endpoints of $f(\lambda)$ on $(0,1)$ and tangent to $f(\lambda)$.

Since

$$f(0) = 1$$

$$f(1) = \tfrac{1}{2}$$

such a line is of the form

$$-\tfrac{1}{2}\lambda + r \tag{4.23}$$

where r is determined by

$$f'(\lambda^*) = -\tfrac{1}{2}\ .$$

Then

$$\lambda^* = \sqrt{2} - 1$$

so

$$-\tfrac{1}{2}\lambda^* + r = -\tfrac{1}{2}$$

and

$$r = \sqrt{2} - \tfrac{1}{2}.$$

Thus

$$f(\lambda) \geq -\tfrac{1}{2}\lambda + (\sqrt{2} - \tfrac{1}{2})$$

for $\lambda \varepsilon (0,1)$. A line analogous to (4.23) which bounds $g(\lambda)$ from below exists but is difficult to determine since we would have to solve the equation

$$\sum_{i=0}^{\infty} i\, \lambda^{i-1} (\rho_i - \rho_{i+1}) = \rho_1$$

for λ. However, we know $(\rho_i - \rho_{i+1}) \geq 0$ for all i, so that

$$g(\lambda) \geq (1-\rho_1) + \lambda(\rho_1-\rho_2)$$

for $\lambda \varepsilon (0,1)$. We have then

$$\begin{aligned} F(\lambda) &= f(\lambda)\; g(\lambda) \\ &\geq [-\frac{\lambda}{2} + (\sqrt{2} - \tfrac{1}{2})][(\rho_1 - \rho_2)\, \lambda + (1 - \rho_1)] \\ &= - \frac{(\rho_1-\rho_2)}{2} \lambda^2 + [\sqrt{2}\, \rho_1 - (\sqrt{2} - \tfrac{1}{2})\, \rho_2 - \tfrac{1}{2}]\lambda \\ &\qquad + (1 - \rho_1)(\sqrt{2} - \tfrac{1}{2}) \\ &= B_\ell(\lambda), \text{ say,} \end{aligned}$$

since both lines are positive on (0,1). The following bounds are thereby obtained for $\Delta(1,\lambda)$:

Upper Bound

$$2\sigma^2(1 - \rho_1)\left(\frac{1}{2(1-\rho_1)}\right)^{\lambda}$$

This bound is actually valid under the conditions of Theorem 4.2 if we put $\rho_1 = \rho_h$. This is because we only need $\Delta(h,\lambda)$ to have a unique minimum in order to derive the bound.

<u>Lower Bound</u>

$$2\sigma^2(a_\ell \lambda^2 + b_\ell \lambda + (h)$$

$$\begin{aligned} B_\ell(0) &= (1 - \rho_1)(\sqrt{2} - \tfrac{1}{2}) \\ &> 0 \\ B_\ell(1) &= (1 - \rho_2)(\sqrt{2} - 1) \\ &> 0 \end{aligned}$$

and $B_\ell(\lambda)$ is a concave parabola so it is a nontrivial lower bound for $\Delta(1,\lambda)$.

To show (4.22) we notice that $\sum_{i=0}^{\infty} i\, \lambda^{i-1}(\rho_i - \rho_{i+2})$ increases from $(\rho_1 - \rho_2)$ to $\sum_{i=1}^{\infty} \rho_i$ on (0,1) and so we must have

$$\begin{aligned} 2\sigma^2(\rho_1 - \rho_2) &\leq \Delta(1,\lambda_0) \\ &\leq 2\sigma^2 \sum_{i=1}^{\infty} \rho_i . \end{aligned}$$

But since

$$\Delta(1,0) = 2\sigma^2(1 - \rho_1)$$

and

$$\begin{aligned} \Delta(1,1) &= 2\sigma^2(\tfrac{1}{2}) \\ &= \sigma^2 \end{aligned}$$

we know that

$$\Delta(1,\lambda_0) \leq \min\{2\sigma^2(1 - \rho_1),\ \sigma^2\} .$$

Also $\sum_{i=1}^{\infty} \rho_i > \tfrac{1}{2}$ by (4.20) so that

$$\sigma^2 < 2\sigma^2 \sum_{i=1}^{\infty} \rho_i$$

and we finally get

$$2\sigma^2(\rho_1 - \rho_2) \leq \Delta(1,\lambda_0) \leq \min\{2\sigma^2(1 - \rho_1),\ \sigma^2\}.$$

This last inequality reveals the fact that the predictor

(4.24) $$\hat{X}_{t+1} = (1 - \lambda_0) \sum_{i=0}^{\infty} \lambda_0^i X_{t-i}$$

is never worse than predicting by using the mean (i.e., a simple unweighted average of the past).

The bounds derived have potential use in cases where λ_0 cannot be explicitly determined but ρ_1 and ρ_2 can be estimated fairly accurately.

Cox-Type Predictors

A natural generalization of the ewma is

(4.25 $$\hat{X}_{t+h} = \frac{1}{f(\lambda)} \sum_{i=0}^{\infty} b_i \lambda^i X_{t-i}$$

where $f(\lambda) = \sum_{i=0}^{\infty} b_i \lambda^i$, in an analytic function in the unit circle, real and nonzero on the interval (-1, +1). Our interest in random variables of this form lies in the fact that they often will be insensitive to long term trends in the mean. By this we mean that if the mean function changes slowly with time the predictor will follow the trend of the mean. We will often be able to modify the predictor and obtain one which is nearly optimal in the sense of smallest mean square error. The following result, however, limits their use.

Theorem 4.3

If a time series obeys the model in (1.1) with M_t a polynomial of degree m then a Cox-type predictor is unbiased for all λ if and only if m = 0 (constant mean).

Proof:

If m = 0 the predictor is clearly unbiased since

$$E\{\frac{1}{f(\lambda)} \sum_{i=0}^{\infty} b_i \lambda^i X_{t-i}\} = \frac{\mu \sum_{i=0}^{\infty} b_i \lambda^i}{f(\lambda)}$$

$$= \mu.$$

If the predictor is unbiased then

$$E\{X_{t+h}\} = M_{t+h}$$

$$= \frac{1}{f(\lambda)} \sum_{i=0}^{\infty} b_i \lambda^i M_{t-i}$$

where M_t is the polynomial mean function of degree m. But we have that

$$M_{t-i} = \sum_{j=0}^{m} \frac{(-i)^j M_t^{(j)}}{j!}$$

where $M_t^{(j)} = d^j M_t/d_t^j$ and

$$M_{t+h} = \sum_{j=0}^{m} \frac{h^j}{j!} M_t^{(j)}$$

so that unbiasedness requires that

$$\sum_{j=0}^{m} \frac{h^j M_t^{(j)}}{j!} = \frac{1}{f(\lambda)} \sum_{i=0}^{\infty} b_i \lambda^i \left(\sum_{j=0}^{m} \frac{(-i)^j M_t^{(j)}}{j!} \right)$$

$$= \frac{1}{f(\lambda)} \sum_{j=0}^{m} \frac{M_t}{j!} \left\{ \sum_{i=0}^{\infty} (-i)^j b_i \lambda^i \right\}$$

or $\sum_{i=0}^{\infty} \frac{(-i)^j b_i \lambda^i}{f(\lambda)} = h^j$ for j = 0, 1, 2, ...

In particular we require

$$\sum_{i=0}^{\infty} \frac{-i\, b_i \lambda^i}{f(\lambda)} = h$$

or

$$\frac{\lambda\, f'(\lambda)}{f(\lambda)} = - h$$

whence

$$f(\lambda) = \lambda^{-h}.$$

This requires that

$$\hat{X}_{t+h} = X_{t+h}$$

which is absurd.

This result on unbiasedness clearly applies when using the ewma to estimate the mean of a process.

When obtaining predictors with arbitrarily small mean square errors we will use the Cox-type expression as estimators of the mean rather than predictors themselves. We will be looking at the performance of these random variables for values of λ in the vicinity of $\lambda = 1$.

Definition

An estimator $\hat{\mu}(\lambda)$ of the constant mean μ of a stationary time series is called Abel-efficient if

$$\lim_{\lambda \to 1-} E\{\hat{\mu}(\lambda) - \mu\}^2 = 0 . \tag{4.26}$$

The following result gives interesting formulae for the variance of an estimate of μ and for the mean square error of prediction of X_{t+h}.

Theorem 4.4

If $\sum_{i=0}^{\infty} |b_i| < \infty$ and $\sum_{i=0}^{\infty} b_i = 1$ for a real sequence of b_i's the random variable

(4.27) $$\sum_{i=0}^{\infty} b_i X_{t-i}$$

has the following properties, provided the mean of the process is a constant μ.

p_1) $$E\{\sum_{i=0}^{\infty} b_i X_{t-i} - \mu\}^2 = \sigma^2 \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} b_i b_j \rho_{|i-j|}$$

(4.28) $$= \sigma^2 \int_{-\pi}^{+\pi} |g(y)|^2 \, d\, F(y)$$

p_2) $$E\{X_{t+h} - \sum_{i=0}^{\infty} b_i X_{t-i}\}^2 = \sigma^2 \Big\{ 1 - 2\sum_{i=0}^{\infty} b_i \rho_{h+i} + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} b_i b_j \rho_{|i-j|} \Big\}$$

$$= \sigma^2 \int_{-\pi}^{+\pi} |g(y) - e^{-ihy}|^2 \, d\, F(y)$$

where $g(y) = \sum_{k=0}^{\infty} b_k e^{iky}$ and $F(y)$ is the spectral distribution function of the process.

Proof:

p_3) $$E\{\sum_{i=0}^{\infty} b_i X_{t-i} - \mu\}^2 = E\{\sum_{i=0}^{\infty} b_i (X_{t-i} - \mu)\}^2$$

$$= E\{\sum_{k=0}^{\infty} \sum_{j=0}^{\infty} b_k b_j (X_{t-k} - \mu)(X_{t-j} - \mu)\}$$

$$= \sigma^2 \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} b_k b_j \rho_{|k-j|} .$$

But we know

$$\rho_{|k-j|} = \int_{-\pi}^{+\pi} e^{i(k-j)y} \, d\, F(y)$$

hence

$$E\{\sum_{i=0}^{\infty} b_i X_{t-i}-\mu\}^2 = \sigma^2 \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} b_k b_j \int_{-\pi}^{+\pi} e^{i(k-j)y} \, d\, F(y)$$

$$= \sigma^2 \int_{-\pi}^{+\pi} (\sum_{k=0}^{\infty} b_k e^{iky})(\sum_{j=0}^{\infty} b_j e^{-ijy}) \, d\, F(y)$$

$$= \sigma^2 \int_{-\pi}^{+\pi} |g(y)|^2 \, d\, F(y).$$

p_2) $$E\{X_{t+h} - \sum_{i=0}^{\infty} b_i X_{t-i}\}^2 = E\{\sum_{i=0}^{\infty} b_i (X_{t+h} - X_{t-i})\}^2$$

$$= \sigma^2 \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} b_i b_j (1 - \rho_{h+i} - \rho_{h+j} + \rho_{|i-j|})$$

$$= \sigma^2 (1 - 2 \sum_{i=0}^{\infty} b_i \rho_{h+i}$$

$$+ \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} b_i b_j \rho_{|i-j|})$$

$$= \sigma^2 (1 - 2 \int_{-\pi}^{+\pi} e^{ihy} g(y) \, d\, F(y)$$

$$+ \int_{-\pi}^{+\pi} |g(y)|^2 \, d\, F(y)) \ .$$

But since F(y) is an even function we know that

$$\int_{-\pi}^{+\pi} e^{ihy} g(y) \, d\, F(y) = \int_{-\pi}^{+\pi} e^{-ihy} \overline{g(y)} \, d\, F(y) \ .$$

Hence

$$E\{\sum_{i=0}^{\infty} b_i X_{t-i} - X_{t+h}\}^2 = \sigma^2 \left\{ \int_{-\pi}^{+\pi} (e^{-ihy} g(y))(e^{ihy} \right.$$

$$\left. - \overline{g(y)}) \, d\, F(y) \right\}$$

$$= \sigma^2 \int_{-\pi}^{+\pi} |e^{ihy} - g(y)|^2 \, d\, F(y).$$

In particular, it is interesting to note that if the b_i are probabilities then the choosing of a distribution on the nonnegative integers which gives good weights for a predictor is equivalent to choosing a characteristic function of such a probability distribution which is "close" to e^{-ihy} in the sense of the norm defined by

$$(4.29) \qquad ||h(y)|| = \sigma\left(\int_{-\pi}^{+\pi} |h(y)|^2 \, d\, F(y)\right)^{\frac{1}{2}}$$

We first look at Cox-type estimators for uncorrelated random variables with common variance. Denote the estimator as

$$(4.30) \qquad \hat{\mu}(\lambda) = (\sum_{i=0}^{\infty} a_i \lambda^i X_{t-i})/f(\lambda)$$

and its variance as $\Delta_{f(\lambda)}$ where $f(\lambda)$ is chosen as in (4.25). It is clear that we may as well confine attention to $\lambda \in (0,1)$ and positive a_i if we want good estimators since

$$\Delta_{f(\lambda)} = \sigma^2 \sum_{i=0}^{\infty} (a_i \lambda^i)^2 / (\sum_{i=0}^{\infty} a_i \lambda^i)^2$$

and

$$\begin{aligned} \Delta_{f(\lambda)} &= |\Delta_{f(\lambda)}| \\ &= \sigma^2 \sum_{i=0}^{\infty} (a_i \lambda^i)^2 / |\sum_{i=0}^{\infty} a_i \lambda^i|^2 \\ &\geq \sigma^2 \sum_{i=0}^{\infty} (a_i \lambda^i)^2 / (\sum_{i=0}^{\infty} |a_i||\lambda|^i)^2. \end{aligned}$$

We also need look only at divergent series $\sum_{i=0}^{\infty} a_i$ since otherwise we cannot have good estimators in the sense of Abel-efficiency.

With this in mind we consider coefficients of the form

$$(4.31) \qquad a_i \sim i^{\nu} L(i)$$

as $i \to \infty$ for $\nu \geq 0$, where $\sim$ means the ratio of the two quantities tends to 1 and $L(i)$ is a nonnegative function of slow growth. This means that

$$L(c\,i)/L(i) \to 1$$

as $i \to \infty$ for all $c > 0$.

Theorem 4.5

If $a_i \sim i^{\nu} L(i)$ then

$$\Delta_{f(\lambda)} \sim \sigma^2 \Gamma(1+2\nu)(1-\lambda)/\Gamma^2(1+\nu)2^{1+2\nu}$$

as $\lambda \to 1-$ for uncorrelated random variables with constant variance σ^2.

Proof:

A result used by Smith [17] and deducible from a Tauberian theorem due to Karamata [8] states that if

$$a_i \sim i^{\nu} L(i)$$

then

$$\sum_{i=0}^{\infty} a_i \lambda^i \sim \frac{\Gamma(1+\nu)}{(1-\lambda)^{1+\nu}} L\left(\frac{1}{1-\lambda}\right)$$

as $\lambda \to 1-$. From this we deduce that since

$$a_i^2 \sim i^{2\nu} L^2(i)$$

and $L^2(i)$ is still a function of slow growth then

$$\sum_{i=0}^{\infty} a_i^2 \lambda^{2i} \sim \frac{\Gamma(1+2\nu)}{(1-\lambda^2)^{1+2\nu}} L^2\left(\frac{1}{1-\lambda^2}\right)$$

and consequently

$$\frac{\Delta_{f(\lambda)}}{\sigma^2} = \sum_{i=0}^{\infty} a_i^2 \lambda^{2i} / (\sum_{i=0}^{\infty} a_i \lambda^i)^2$$

$$\sim \frac{\Gamma(1+2\nu)}{(1-\lambda^2)^{1+2\nu}} L^2(\frac{1}{1-\lambda^2}) \frac{(1-\lambda)^{2(1+\nu)}}{\Gamma^2(1+\nu)} \{L(\frac{1}{1-\lambda})\}^{-2}.$$

Now

$$L^2(\frac{1}{1-\lambda^2}) = L^2(\frac{1}{(1-\lambda)(1+\lambda)})$$

$$\sim L^2(\frac{1}{2(1-\lambda)})$$

therefore

$$L^2(\frac{1}{1-\lambda^2})\{L(\frac{1}{1-\lambda})\}^{-2} \sim 1$$

as $\lambda \to 1-$ and the function of slow growth does not affect the final result. Also

$$(1-\lambda)^{2+2\nu}/(1-\lambda^2)^{1+2\nu} = (1-\lambda)^{2+2\nu}/(1-\lambda)^{1+2\nu}(1+\lambda)^{1+2\nu}$$

$$\sim (1-\lambda)/2^{1+2\nu}$$

hence

$$\Delta_{f(\lambda)} \sim \frac{\sigma^2\Gamma(1+2\nu)(1-\lambda)}{\Gamma^2(1+\nu)2^{1+2\nu}}$$

as $\lambda \to 1-$. If ν is also large we have by Stirling's formula

$$\frac{\Gamma(1+2\nu)}{\Gamma^2(1+\nu)2^{1+2\nu}} \sim \frac{1}{2\sqrt{\pi\nu}}$$

so for large ν and λ near 1 we have

$$\Delta_{f(\lambda)} \approx \frac{\sigma^2(1-\lambda)}{2\sqrt{\pi\nu}} .$$

The next result shows we have quite a wide choice of functions on which to base our estimator in the case of a stationary sequence of random variables.

Theorem 4.6

Suppose $f(z)$ is analytic in and on the unit circle except at $z = 1$ where it has a pole of order k and k is a positive integer. Suppose $f(z) \neq 0$ for $z \in (0,1)$ and for $|z| = 1$, then the Cox estimator of μ based on $f(z)$ is Abel-efficient if and only if the spectral distribution function, $F(y)$, of the process is continuous at $y = 0$.

Proof:

The Cox estimator based on $f(z)$ is

$$\hat{\mu}(\lambda) = \frac{1}{f(\lambda)} \sum_{i=0}^{\infty} a_i \lambda^i X_{t-i}$$

where $\sum_{i=0}^{\infty} a_i z^i$ is the Taylor expansion of $f(z)$ in the unit circle. If in theorem 4.4 we put

$$b_i = a_i \lambda^i / f(\lambda)$$

then
$$g(y) = \frac{1}{f(\lambda)} \sum_{k=0}^{\infty} a_k \lambda^k e^{iky}$$

$$= f(\lambda e^{iy})/f(\lambda)$$

and consequently

$$\text{Var. } (\hat{\mu}(\lambda)) = \sigma^2 \int_{-\pi}^{+\pi} \frac{|f(\lambda e^{iy})|^2}{|f(\lambda)|^2} \, d F(y).$$

$f(z)$ may be written as

$$f(z) = h(z)/(1-z)^k$$

where $h(z)$ is analytic in and on the unit circle. Then

$$\left|\frac{f(\lambda e^{iy})}{f(\lambda)}\right|^2 = \frac{|1-\lambda|^{2k}}{|1-\lambda e^{iy}|^{2k}} \frac{|h(\lambda e^{iy})|^2}{|h(\lambda)|^2}$$

and the analyticity of $h(z)$ for $|z| \leq 1$ implies $|h(\lambda\ e^{iy})|^2/|h(\lambda)|^2$ is bounded uniformly for $\lambda\ \varepsilon(0,1)$ and $y\ \varepsilon(-\pi,+\pi)$. The quantity is also bounded away from zero for $y\ \varepsilon(-\pi,+\pi)$ and $\lambda\ \varepsilon(1-\gamma,1)$ for some $\gamma > 0$.

Since

$$|(1-\lambda)|/|1-\lambda\ e^{iy}| < 1$$

then by bounded convergence we have

$$\lim_{\lambda\to 1-} \operatorname{Var}(\hat{\mu}(\lambda)) = \sigma^2 \int_{-\pi}^{+\pi} \lim_{\lambda\to 1-} \frac{|f(\lambda\ e^{iy})|^2}{|f(\lambda)|^2}\ d\ F(y)$$

so

$$\sigma^2\ m \int_{-\pi}^{+\pi} \delta(y)\ d\ F(y) \leq \lim_{\lambda\to 1-} \operatorname{Var}(\hat{\mu}(\lambda)) \leq \sigma^2\ M \int_{-\pi}^{+\pi} \delta(y)\ d\ F(y)$$

where

$$M = \sup_{\substack{\lambda\varepsilon(0,1)\\ y\varepsilon(-\pi,+\pi)}} \left|\frac{h(\lambda\ e^{iy})}{h(\lambda)}\right|^2$$

$$m = \inf_{\substack{\lambda\varepsilon(1-\gamma,1)\\ y\varepsilon(-\pi,+\pi)}} \left|\frac{h(\lambda\ e^{iy})}{h(\lambda)}\right|^2$$

$$\delta(y) = \begin{cases} 0 & \text{if} \quad y \neq 0 \\ 1 & \text{if} \quad y = 0 \end{cases}.$$

Since $m > 0$, the result follows.

The next result concerns the particular function $f(z) = 1/1-z$. We know already that the Cox estimator based on this function is generally Abel-efficient. The following expresses the requirements in terms of the correlation function.

Theorem 4.7

If $f(z) = 1/1-z$ then the Cox estimator based on $f(z)$ is Abel-efficient if and only if

$$\sum_{r=1}^{\infty} \lambda^r \rho_r = o(\frac{1}{1-\lambda})$$

as $\lambda \to 1-$.

Proof:

$$\hat{\mu}(\lambda) = (1-\lambda) \sum_{i=0}^{\infty} \lambda^i X_{t-i}$$

$$\begin{aligned} \mathrm{Var}(\hat{\mu}(\lambda)) &= \sigma^2(1-\lambda)^2 \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \lambda^{i+j} \rho_{|i-j|} \\ &= \frac{\sigma^2(1-\lambda)}{(1+\lambda)}\{1 + 2 \sum_{r=1}^{\infty} \lambda^2 \rho_r\} \\ &\sim \sigma^2(1-\lambda)\{\sum_{r=1}^{\infty} \lambda^r \rho_r + \tfrac{1}{2}\} \end{aligned}$$

as $\lambda \to 1-$ and the result follows.

In view of theorem 4.5 for uncorrelated variables, we intuitively feel that perhaps for any given Cox-type estimator we can improve on it by choosing an $f(z)$ whose coefficient sequence increases more rapidly than the given one. In other words, it seems that if $f(z)$ is of the form $g(z)/(1-z)^k$ then we get better and better estimators of μ by taking k larger and larger. The following example lends some support to this intuition.

Lemma 4.3

Let $\hat{\mu}_1(\lambda)$ be the Cox-type estimator of μ based on $f(z) = 1/1-z$ and let $\hat{\mu}_2(\lambda)$ be the estimator based on $z/(1-z)^2$. If the process is first order autoregressive with

$$\rho_\tau = \rho^\tau , \qquad \tau = 0, 1, 2, \ldots$$

then

$$\mathrm{Var}(\hat{\mu}_1(\lambda)) = \frac{\sigma^2(1-\lambda)}{(1+\lambda)} \left\{\frac{1+\lambda\rho}{1-\lambda\rho}\right\}$$

(4.32)

$$\mathrm{Var}(\hat{\mu}_2(\lambda)) = \frac{\sigma^2(1-\lambda)^2}{(1+\lambda)^2} \left\{\frac{(1+\lambda^2)(1+\lambda\rho)}{(1-\lambda^2)(1-\lambda\rho)} + \frac{2\lambda\rho}{(1-\lambda\rho)^2}\right\}$$

and

$$(4.33) \quad \mathrm{Var}(\hat{\mu}_1(\lambda)) > \mathrm{Var}(\hat{\mu}_2(\lambda)) , \qquad \lambda \in (0,1).$$

Proof:

Computation of the variances is routine so we will only prove (4.33). We start by noting that since $\lambda \in (0,1)$ then

$$\frac{2\lambda}{1-\lambda^2} > \frac{2\lambda\rho}{1-(\lambda\rho)^2}$$

and

$$\frac{1}{1-\lambda} \left(\frac{(1+\lambda)^2-(1+\lambda^2)}{1+\lambda}\right) > \frac{2\lambda\rho}{1-(\lambda\rho)^2}$$

or

$$\frac{1+\lambda}{1-\lambda} - \frac{1+\lambda^2}{1-\lambda^2} > \frac{2\lambda\rho}{(1-\lambda\rho)(1+\lambda\rho)} .$$

Hence

$$\frac{1+\lambda}{1-\lambda} > \frac{1+\lambda^2}{1-\lambda^2} + \frac{2\lambda\rho}{(1-\lambda\rho)(1+\lambda\rho)}$$

and

$$\frac{\sigma^2(1-\lambda)}{(1+\lambda)} \left\{\frac{1+\lambda\rho}{1-\lambda\rho}\right\} > \frac{\sigma^2(1-\lambda)^2}{(1+\lambda)^2} \left\{\frac{(1+\lambda^2)(1+\lambda\rho)}{(1-\lambda^2)(1-\lambda\rho)} + \frac{2\lambda\rho}{(1-\lambda\rho)^2}\right\}.$$

Since Cox-type expressions may be used to predict future values of the process as well as estimate the mean, one naturally wonders if perhaps a relation like (4.33) holds for mean square errors of prediction as well as for variances. The following results, which compare the mean square errors for the process in Lemma 4.3, gives a negative answer to this question.

Lemma 4.4

Let $\hat{X}_{1,t+h}$ be the Cox-type predictor of X_{t+h} based on $f(z) = 1/1-z$ and let $\hat{X}_{2,t+h}$ be the predictor based on $z/(1-z)^2$. If the process is first order autoregressive with

$$\rho_\tau = \rho^\tau , \qquad \tau = 0, 1, 2, \ldots$$

then

$$\text{(4.34)} \quad \text{i)} \quad E\{X_{t+h} - \hat{X}_{1,t+h}\}^2 = \frac{2\sigma^2}{1+\lambda}\left(1 + \frac{(1-\lambda)(\lambda\rho-\rho^h-\lambda\rho^h)}{(1-\lambda\rho)}\right)$$

$$\text{ii)} \quad E\{X_{t+h} - \hat{X}_{2,t+h}\}^2 = \sigma^2\left\{1 - \frac{2\rho^h(1-\lambda)^2}{(1-\lambda\rho)^2} + \frac{(1+\lambda^2)(1+\lambda\rho)(1-\lambda)}{(1+\lambda)^3(1-\lambda\rho)} + \frac{2\lambda\rho(1-\lambda)^2}{(1+\lambda)^2(1-\lambda\rho)^2}\right\}.$$

Proof:

The first of these formulae has been given earlier in this work and we only derive the second one here. Since X_{t+h} and $\hat{X}_{2,t+h}$ have the same expectation we have

$$E\{X_{t+h} - \hat{X}_{2,t+h}\}^2 = \mathrm{Var}(X_{t+h}) + \mathrm{Var}(\hat{X}_{2,t+h}) - 2\,\mathrm{Cov}(X_{t+h},\hat{X}_{2,t+h}).$$

From lemma 4.3 we have the variance of $\hat{X}_{2,t+h}$ so we need only compute $\mathrm{Cov}(X_{t+h},\hat{X}_{2,t+h})$.

We write

$$\hat{X}_{2,t+h} = (1-\lambda)^2 \sum_{i=0}^{\infty} i\,\lambda^{i-1}\,X_{t+1-i}$$

so that

$$\mathrm{Cov}(\hat{X}_{2,t+h},X_{t+h}) = \sigma^2(1-\lambda)^2 \sum_{i=0}^{\infty} i\,\lambda^{i-1}\,\rho^{h+i-1}$$

$$= \frac{\sigma^2(1-\lambda)^2\rho^h}{(1-\lambda\rho)^2}.$$

Then

$$E\{X_{t+h} - \hat{X}_{2,t+h}\}^2 = \sigma^2\left\{1 - \frac{2\rho^h(1-\lambda)^2}{(1-\lambda\rho)^2} + \frac{(1+\lambda^2)(1+\lambda\rho)(1-\lambda)}{(1+\lambda)^3(1-\lambda\rho)} + \frac{2\lambda\rho(1-\lambda)^2}{(1+\lambda)^2(1-\lambda\rho)^2}\right\}.$$

Whereas we have seen from lemma 4.3 that

$$(4.35) \qquad \mathrm{Var}(\hat{X}_{2,t+h}) < \mathrm{Var}(\hat{X}_{1,t+h}),\ \lambda \in (0,1)$$

we also see that

$$(4.36) \qquad \frac{(1-\lambda)^2\rho^h}{(1-\lambda\rho)^2} < \frac{\rho^h(1-\lambda)}{(1-\lambda\rho)}$$

for $\rho^h > 0$. The expression on the right of (4.36) is $\mathrm{Cov}(X_{t+h}\ \hat{X}_{1,t+h})$ so we see that the relation between [4.34 (i)] and [4.34 (ii)] is rather complicated, depending on whether ρ^h

is positive or negative. We can conclude from (4.35) and (4.36) that

$$E\{X_{t+h} - \hat{X}_{2,t+h}\}^2 < E\{X_{t+h} - \hat{X}_{1,t+h}\}^2 \tag{4.37}$$

if $\rho^h < 0$. Below is a tabular comparison of these two predictors for cases where ρ^h is not necessarily negative.

Let

$$\Delta_i(\lambda,h,\rho) = E\{X_{t+h} - \hat{X}_{i,t+h}\}^2$$

for $i = 1, 2$. These mean square errors will be compared for values of ρ between $-.9$ and $.9$. For each value of ρ in this range the following quantity is tabled for $h = 1, 2$:

$$\Delta_i(h,\rho) = \min_{\lambda \varepsilon S} \frac{\Delta_i(\lambda,h,\rho)}{\sigma^2}$$

where $S = \{-.9, -.8, \ldots, .8, .9\}$.

We now show how an almost optimal linear predictor may be constructed when the mean of the process is not known, using the best linear predictor for known mean and a Cox-type estimator of the mean.

<u>Theorem 4.8</u>

Suppose the best "linear" predictor of X_{t+h} when the constant mean μ of the process is known is

$$\sum_{i=0}^{\infty} a_i X_{t-i} + \mu(1 - \sum_{i=0}^{\infty} a_i). \tag{4.38}$$

Then

ρ	-.9	-.8	-.7	-.6	-.5	-.4	-.3	-.2	-.1	.1	.2	.3	.4	.5	.6	.7	.8	.9
$\Delta_1(1,\rho)$	1.10	1.10	1.10	1.09	1.09	1.08	1.08	1.07	1.06	1.04	1.03	1.01	.98	.89	.75	.58	.40	.20
$\Delta_2(1,\rho)$	1.00	1.01	1.01	1.01	1.01	1.02	1.02	1.02	1.02	1.03	1.03	1.04	1.04	.91	.60	.58	.40	.20
$\Delta_1(2,\rho)$	.35	.60	.78	.90	.97	1.00	1.02	1.03	1.04	1.06	1.07	1.07	1.06	1.05	1.02	.94	.70	.38
$\Delta_2(2,\rho)$	.36	.60	.78	.91	.99	1.01	1.01	1.02	1.02	1.03	1.04	1.05	1.05	1.06	1.06	.96	.70	.38

(4.39) $$\hat{X}_{t+h} = \sum_{i=0}^{\infty} a_i X_{t-i} + \hat{\mu}(1 - \sum_{i=0}^{\infty} a_i)$$

where

$$\hat{\mu} = \sum_{i=0}^{\infty} b_i X_{t-i}$$

has mean square error

$$E\{X_{t+h} - \hat{X}_{t+h}\}^2 = \sigma_0^2 + (1 - \sum_{i=0}^{\infty} a_i)^2 E\{\hat{\mu} - \mu\}^2$$

where

(4.40) $$\sigma_0^2 = \lim_{n\to\infty} \sigma^2(1 - \underline{\rho}' \Sigma^{-1} \underline{\rho})$$

is the minimum mean square error of prediction for known mean (i.e., mean square error of 4.38).

<u>Proof:</u>

$$E\{\hat{X}_{t+h} - X_{t+h}\}^2 = E\{X_{t+h} - \sum_{i=0}^{\infty} a_i X_{t-i} - \hat{\mu}(1 - \sum_{i=0}^{\infty} a_i)\}^2$$

$$= E\{(X_{t+h}-\mu) - \sum_{i=0}^{\infty} a_i (X_{t-i}-\mu)$$

$$- (\hat{\mu}-\mu)(1 - \sum_{i=0}^{\infty} a_i)\}^2$$

$$= E\{(X_{t+h}-\mu) - \sum_{i=0}^{\infty} a_i(X_{t-i}-\mu)\}^2$$

$$+ (1 - \sum_{i=0}^{\infty} a_i)^2 E\{\hat{\mu}-\mu\}^2$$

$$- 2(1 - \sum_{i=0}^{\infty} a_i) E\{(X_{t+h}-\mu - \sum_{i=0}^{\infty} a_i(X_{t-i}-\mu))$$

$$(\hat{\mu}-\mu)\}.$$

Now $(X_{t+h} - \mu - \sum_{i=0}^{\infty} a_i(X_{t-i}-\mu))$ is recognized as the error of prediction of the best linear predictor when the mean is known and hence is uncorrelated with each member of the sequence $(X_t-\mu)$, $(X_{t-i}-\mu)$, ... In particular, the error is uncorrelated with $(\hat{\mu}-\mu)$ since

$$(\hat{\mu}-\mu) = \sum_{i=0}^{\infty} b_i(X_{t-i}-\mu) + (\sum_{i=0}^{\infty} b_i-1)\mu.$$

We see then that since

$$E\{(X_{t+h}-\mu) - \sum_{i=0}^{\infty} a_i(X_{t-i}-\mu)\}^2 = \sigma_0^2$$

then

$$E\{\hat{X}_{t+h} - X_{t+h}\}^2 = \sigma_0^2 + (1 - \sum_{i=0}^{\infty} a_i)\ E\{\hat{\mu}-\mu\}^2.$$

If an Abel-efficient estimator is used for $\hat{\mu}$ then we get

$$\lim_{\lambda\to 1-} E\{\hat{X}_{t+h} - X_{t+h}\}^2 = \sigma_0^2.$$

It will often happen (e.g., with autoregressive schemes) that in constructing predictors as described above, only a finite number of the a_i's will be nonzero and in this case we have

$$\hat{X}_{t+h} = \sum_{i=0}^{k} a_i\ X_{t-i} + \hat{\mu}(1 - \sum_{i=0}^{k} a_i). \tag{4.41}$$

If $\hat{\mu}$ is chosen to be insensitive to long term trends (i.e., $\hat{\mu} = (1-\lambda) \sum_{i=0}^{\infty} \lambda^i X_{t-i}$, λ not too near 1) then the predictor will be insensitive to such trends if k is not too large. This will be the case for low order autoregressive schemes.

<u>Example 4.1</u>

Suppose we are dealing with the simplest autoregressive scheme. The classical predictor of X_{t+h} is

$$\rho^h\ X_t + (1-\rho^h)\mu$$

and putting

$$\hat{\mu}(\lambda) = (1-\lambda) \sum_{i=0}^{\infty} \lambda^i X_{t-i}$$

the modified Cox predictor is

$$(1-\lambda+\lambda\rho^h) X_t + (1-\rho^h)(1-\lambda) \sum_{i=1}^{\infty} \lambda^i X_{t-i}.$$

Its mean square error from theorem 4.8 is

$$E\{\hat{X}_{t+h} - X_{t+h}\}^2 = \sigma_0^2 + (1-\rho^h)^2 \operatorname{Var}\{(1-\lambda) \sum_{i=0}^{\infty} \lambda^i X_{t-i}\}$$

$$= \sigma^2(1-\rho^{2h}) + \frac{\sigma^2(1-\rho^h)^2(1-\lambda)^2}{(1-\lambda^2)(1-\lambda\rho)}$$

$$= \sigma^2(1 - \rho^{2h} + \frac{(1-\rho^h)^2(1-\lambda)}{(1+\lambda)(1-\lambda\rho)}).$$

It should be mentioned that one cannot decrease the mean square error at will and at the same time guard against trends in the mean. One must in each experimental situation decide which type of error is most important and choose λ accordingly. One point, however, is quite consoling. It may happen (see Cox [3]) that the mean square error of prediction is rather insensitive to choices of λ. In this case one may well choose λ considerably below 1 and still enjoy a small mean square error while maintaining a guard against long term trends.

We have seen that the predictor in (4.39), where $\hat{\mu}$ is an Abel-efficient estimator of the mean μ, is optimal as $\lambda \to 1-$. We have also noted in (4.41) that (4.30) is often of the form

$$\hat{X}_{t+h} = \sum_{i=0}^{k-1} a_i X_{t-i} + a_k \hat{\mu} \tag{4.42}$$

where

$$\sum_{i=0}^{k} a_i = 1$$

for some positive finite integer k. Now whereas the a_i's in (4.42) are good choices for λ near 1, it is very unlikely that they are the best choices for all λ. We consider the problem of determining the best a_i's when $\hat{\mu}$ is chosen as in theorem 4.6.

Theorem 4.9

Suppose for a process satisfying the model in (1.1) that

$$\hat{X}_{t+h} = \sum_{i=0}^{k-1} a_i X_{t-i} + a_k \frac{\sum_{i=0}^{\infty} b_i \lambda^i X_{t-r-i}}{f(\lambda)} \tag{4.43}$$

for some positive integers k and r and fixed function

$$f(\lambda) = \sum_{i=0}^{\infty} b_i \lambda^i$$

satisfying the conditions of theorem 4.6. If the mean function M_t is a polynomial of degree $m < k$, then the best choices of the a_i's in (4.43) subject to the predictor being unbiased are

$$\tilde{\underline{a}}_k = \Sigma_k^{-1} \underline{\rho}_k - \Sigma_k^{-1} M_k (M_k' \Sigma_k^{-1} M_k)^{-1} (M_k' \Sigma_k^{-1} \underline{\rho}_k - \underline{h}) \tag{4.44}$$

and the corresponding mean square error is

$$\sigma^2\{1 - \underline{\rho}_k' \Sigma_k^{-1} \underline{\rho}_k + (M_k' \Sigma_k^{-1} \underline{\rho}_k - \underline{h})' (M_k' \Sigma_k^{-1} M_k)^{-1} (M_k' \Sigma_k^{-1} \underline{\rho}_k - \underline{h})\} \tag{4.45}$$

where

$$\tilde{\underline{a}}_k' = (\tilde{a}_0, \tilde{a}_1, \ldots, \tilde{a}_k)$$

$$\underset{(k+1)\times(k+1)}{\Sigma_k} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} & g_r \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} & g_{r-1} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \rho_{k-1} & \rho_{k-2} & \cdots & & 1 & g_{r-k+1} \\ g_r & g_{r-1} & \cdots & & g_{r-k+1} & g_{00} \end{bmatrix}$$

$$\underline{\rho}_k' = (\rho_h, \rho_{h+1}, \ldots, \rho_{h+k-1}, g_{h+r})$$

$$\underset{(m+1)\times(k+1)}{M_k'} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & b_{0r} \\ 0 & 1 & 2 & \cdots & k-1 & b_{1r} \\ 0 & 1 & 2 & \cdots & (k-1)^2 & b_{2r} \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 1 & 2^m & & (k-1)^m & b_{mr} \end{bmatrix},$$

$$\underline{h}' = (1, -h, \ldots, (-h)^m)$$

and

$$g_{00} = [f(\lambda)]^{-2} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} b_i\, b_j\, \lambda^{i+j} \rho_{|i-j|}$$

$$g_{r-j} = [f(\lambda)]^{-1} \sum_{i=0}^{\infty} b_i\, \lambda^i \rho_{|i+r-j|}$$

$$b_{jr} = [f(\lambda)]^{-1} \sum_{i=0}^{\infty} b_i\, \lambda^i (r+i)^j .$$

If m = 0 (constant mean), $\sum_{i=0}^{\infty} b_i \rho_{|i+r-j|}$ is convergent for $j = 0, 1, \ldots k-1, -h$ and the spectral distribution function of the process has a positive continuous density in a neighborhood of the origin, then the expression in (4.45) tends to

(4.46) $$\sigma^2\{1 - \underline{\rho}'_{k-1} \Sigma_{k-1}^{-1} \underline{\rho}_{k-1}\}$$

as $\lambda \to 1-$, where

$$\Sigma_{k-1} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_2 & \cdots & \rho_{k-2} \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \rho_{k-1} & \rho_{k-2}, & \cdots & & 1 \end{pmatrix}$$

$$\underline{\rho}'_{k-1} = (\rho_h, \rho_{h+1}, \cdots \rho_{h+k-1}).$$

Proof:

Let

$$Z_{t-i} = X_{t-i}, \qquad i = 0, 1, 2, \ldots, k-1$$

$$Z_{t-k} = [f(\lambda)]^{-1} \sum_{i=0}^{\infty} b_i \lambda^i X_{t-r-i}$$

and put

$$\underline{Z}' = (Z_t, Z_{t-1}, \ldots, Z_{t-k}).$$

Then

$$\text{Cov.}(\underline{Z}) = \sigma^2 \Sigma_k.$$

The predictor in (4.43) may then be written as

$$\hat{X}_{t+h} = \underline{a}' \underline{Z}$$

where

$$\underline{a}' = (a_0, a_1, \ldots, a_k).$$

For unbiasedness we require that

$$E(\hat{X}_{t+h}) = \sum_{i=0}^{k-1} a_i M_{t-i} + \frac{a_k}{f(\lambda)} \sum_{i=0}^{\infty} b_i \lambda^i M_{t-r-i}$$

$$= M_{t+h}.$$

Since

$$M_{t+r} = \sum_{i=0}^{m} \frac{r^i M_t^{(i)}}{i!}$$

where $M_t^{(i)}$ is the i-th derivative of M_t, then we require that

$$\sum_{i=0}^{k-1} a_i \left(\sum_{j=0}^{m} \frac{(-i)^j M_t^{(j)}}{j!} \right) + \frac{a_k}{f(\lambda)} \sum_{i=0}^{\infty} b_i \lambda^i \left(\sum_{j=0}^{m} \frac{(-r-i)^j}{j!} M_t^{(j)} \right) = \sum_{j=0}^{m} \frac{h^j}{j!} M_t^{(j)}$$

or

$$(4.47) \quad \left\{ \begin{array}{l} \sum_{i=0}^{k} a_i = 1 \\ \sum_{i=0}^{k-1} i\, a_i + \frac{a_k}{f(\lambda)} \sum_{i=0}^{\infty} (r+i)\, b_i \lambda^i = -h \\ \quad\vdots \\ \sum_{i=0}^{k-1} i^m a_i + \frac{a_k}{f(\lambda)} \sum_{i=0}^{\infty} (r+i)^m b_i \lambda^i = (-h)^m \end{array} \right.$$

which may be written as

$$(4.48) \qquad M_k \underline{a} = \underline{h}.$$

The mean square error of prediction of (4.43) for any set of a_i's which make the predictor unbiased is

$$(4.49) \qquad \sigma^2\{1-2\underline{a}'\underline{\rho}_k+\underline{\rho}_k' \Sigma_k \underline{\rho}_k\}.$$

We now see that the minimization of (4.49) subject to (4.48) is exactly the same problem dealt with in theorem 3.1, the minimization of (3.4) subject to (3.5). Hence the solution is given by (3.1) and the mean square error by (3.2) which are (4.44) and (4.45) respectively in the present case.

If m = 0 then (4.45) reduces to

$$(4.50) \qquad \sigma^2\left\{1 - \underline{\rho}_k' \Sigma_k^{-1} \underline{\rho}_k + \frac{(\underline{i}' \Sigma_k^{-1} \underline{\rho}_k - 1)^2}{\underline{i}' \Sigma_k^{-1} \underline{i}}\right\}$$

since

$$M_k = \underline{i}$$

and

$$\underline{h} = 1.$$

By lemma 3.1 with

$$\begin{aligned} \underline{x} &= \underline{y} \\ &= \underline{\rho}_k \end{aligned}$$

and

$$S = \Sigma_k$$

we have

$$\underline{\rho}_k' \Sigma_k^{-1} \underline{\rho}_k = \underline{\rho}_{k-1}' \Sigma_{k-1}^{-1} \underline{\rho}_{k-1} + \frac{(g_{h+r} - \underline{g}' \Sigma_{k-1}^{-1} \underline{\rho}_{k-1})^2}{(g_{00} - \underline{g}' \Sigma_{k-1}^{-1} \underline{g})}$$

where $\qquad \underline{g} = (g_r, g_{r-1}, \ldots, g_{r-k+1})$.

Hence (4.50) becomes

$$(4.51) \qquad \sigma^2\left\{1 - \underline{\rho}_{k-1}' \Sigma_{k-1}^{-1} \underline{\rho}_{k-1} - \frac{(g_{h+r} - \underline{g}' \Sigma_{k-1}^{-1} \underline{\rho}_{k-1})^2}{(g_{00} - \underline{g}' \Sigma_{k-1}^{-1} \underline{g})} + \frac{(\underline{i}' \Sigma_k^{-1} \underline{\rho}_k - 1)^2}{\underline{i}' \Sigma_k^{-1} \underline{i}}\right\}.$$

Now

$$\frac{(g_{h+r}-\underline{g}'\ \Sigma_{k-1}^{-1}\ \underline{\rho}_{k-1})^2}{(g_{00}-\underline{g}'\Sigma_{k-1}^{-1}\ \underline{g})} = \frac{(\ell_{h+r}-\underline{\ell}'\ \Sigma_{k-1}^{-1}\ \underline{\rho}_{k-1})^2}{(\ell_{00}-\underline{\ell}'\ \Sigma_{k-1}^{-1}\ \underline{\ell})}$$

where $\ell_j = f(\lambda)\ g_j$, $\ell_{00} = f^2(\lambda)\ g_{00}$, $\underline{\ell}' = (\ell_r,\ \ell_{r-1},\ \ldots,\ \ell_{r-k+1})$.

Then

$$\lim_{\lambda\to 1-} \ell_j = \lim_{\lambda\to 1-} \sum_{i=0}^{\infty} b_i\ \lambda^i\ \rho_{|i+j|}$$

$$= \sum_{i=0}^{\infty} b_i\ \rho_{|i+j|}$$

$$= p_j,\ \text{say},$$

for $j = r, r-1, \ldots, r-k+1$ and $h+r$. Since Σ_{k-1}^{-1} and $\underline{\rho}_{k-1}$ do not depend on λ then

$$\lim_{\lambda\to 1-} \underline{\ell}'\ \Sigma_{k-1}^{-1}\ \underline{\rho}_{k-1} = \underline{p}'\ \Sigma_{k-1}^{-1}\ \underline{\rho}_{k-1}$$

and

$$\lim_{\lambda\to 1-} \underline{\ell}'\ \Sigma_{k-1}^{-1}\ \underline{\ell} = \underline{p}'\ \Sigma_{k-1}^{-1}\ \underline{p}$$

where $\underline{p}' = (p_r,\ p_{r-1},\ \ldots,\ p_{r-k+1})$. Now

$$\lim_{\lambda\to 1-} \ell_{00} = \lim_{\lambda\to 1-} \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} b_i\ b_j\ \lambda^{i+j}\ \rho_{|i-j|}$$

$$= \lim_{\lambda\to 1-} \int_{-\pi}^{+\pi} |f(\lambda e^{iy})|^2\ d\ G(y)$$

where $G(y)$ is the spectral distribution of the process (see theorem 4.4). Let the density of $G(y)$ be $g(y)$, then by Fatou's lemma we have

$$\underline{\lim}_{\lambda\to 1-} \int_{-\pi}^{+\pi} |f(\lambda e^{iy})|^2\ g(y)\ d\ y \geq \int_{-\pi}^{+\pi} \underline{\lim}_{\lambda\to 1-} |f(\lambda e^{iy})|^2\ g(y)\ d\ y$$

$$= \infty.$$

Hence
$$\lim_{\lambda\to 1-} \ell_{00} = \infty$$

and

$$(4.52)\quad \lim_{\lambda\to 1-} \frac{(g_{h+r}-\underline{g}'\ \Sigma_{k-1}^{-1}\ \underline{\rho}_{k-1})^2}{(g_{00}-\underline{g}'\ \Sigma_{k-1}^{-1}\ \underline{g})} = \frac{(p_{h+r}-\underline{p}'\ \Sigma_{k-1}^{-1}\ \underline{\rho}_{k-1})^2}{(\lim_{\lambda\to 1-} \ell_{00}-\underline{p}'\ \Sigma_{k-1}^{-1}\ \underline{p})} = 0.$$

Using lemma 3.1 again we may write

$$\underline{j}'\ \Sigma_k^{-1}\ \underline{\rho}_k = \underline{j}'\ \Sigma_{k-1}^{-1}\ \underline{\rho}_{k-1} + \frac{(1-\underline{g}'\ \Sigma_{k-1}^{-1}\ \underline{j}\,)(g_{h+r}-\underline{g}'\ \Sigma_{k-1}^{-1}\ \underline{\rho}_{k-1})}{(g_{00}-\underline{g}'\ \Sigma_{k-1}^{-1}\ \underline{g})}$$

and

$$\underline{j}'\ \Sigma_k^{-1}\ \underline{j} = \underline{j}'\ \Sigma_{k-1}^{-1}\ \underline{j} + \frac{(1-\underline{g}'\ \Sigma_{k-1}^{-1}\ \underline{j})^2}{(g_{00}-\underline{g}'\ \Sigma_{k-1}^{-1}\underline{g})}.$$

We see then that

$$\frac{(\underline{j}'\ \Sigma_k^{-1}\ \ \underline{\rho}_k - 1)^2}{\underline{j}'\Sigma_k^{-1}\ \underline{j}} = \frac{C_\lambda A^2 + 2AR_\lambda S_\lambda + R_\lambda^2 S_\lambda/C_\lambda}{BC_\lambda + R_\lambda^2}$$

where

$$A = (\underline{j}'\ \Sigma_{k-1}^{-1}\ \underline{\rho}_{k-1} - 1)$$

$$B = \underline{j}'\ \Sigma_{k-1}^{-1}\ \underline{j}$$

$$C_\lambda = g_{00}-\underline{g}'\ \Sigma_{k-1}^{-1}\ \underline{g}$$

$$R_\lambda = (1-\underline{g}'\ \Sigma_{k-1}^{-1}\ \underline{j})$$

$$S_\lambda = (g_{h+r}-\underline{g}'\ \Sigma_{k-1}^{-1}\ \underline{\rho}_{k-1}).$$

A and B are independent of λ and we note that

$$\lim_{\lambda\to 1-} C_\lambda = 0$$

from theorem 4.6.

Also

$$\lim_{\lambda \to 1-} R_\lambda = 1$$

$$\lim_{\lambda \to 1-} S_\lambda = 0$$

since the components of $\underline{g}$ and g_{h+r} tend to zero as $\lambda \to 1-$. Then we have

$$\lim_{\lambda \to 1-} \frac{(\underline{j}' \Sigma_k^{-1} \underline{\rho}_k - 1)^2}{\underline{j}' \Sigma_k^{-1} \underline{j}} = \lim_{\lambda \to 1-} \frac{R_\lambda^2 S_\lambda^2}{C_\lambda (BC_\lambda + R_\lambda^2)}$$

$$= \lim_{\lambda \to 1-} \frac{S_\lambda^2}{C_\lambda (BC_\lambda / R_\lambda^2 + 1)}$$

$$= \lim_{\lambda \to 1-} \frac{S_\lambda^2}{C_\lambda}$$

$$= \lim_{\lambda \to 1-} \frac{(g_{h+r} - \underline{g}' \Sigma_{k-1}^{-1} \underline{\rho}_{k-1})^2}{(g_{00} - \underline{g}' \Sigma_{k-1}^{-1} \underline{g})}$$

$$= 0$$

from (4.52). Then (4.51) tends to (4.46) as $\lambda \to 1-$.

The expression in (4.46) is recognized as the mean square error of prediction of the best linear unbiased predictor based on the last k observations when the mean is known. Since for an autoregressive scheme of order $\ell \leq k$ the best linear predictor for known mean never depends on more than the last ℓ observations we deduce that the predictor in (4.43) with the a_i's given by (4.44) is optimal as $\lambda \to 1-$ for such schemes when the mean of the process is a constant.

Example 4.2

Let

$$m = 0$$

$$r = 0$$

$$k = 1$$

$$f(\lambda) = 1/1-\lambda$$

then (4.43) becomes

$$\hat{X}_{t+h} = a_0 \; X_t + a_1 (1-\lambda) \sum_{i=0}^{\infty} \lambda^i \; X_{t-i}$$

and this is of the form of the modified Cox predictor in example 4.1.

Since

$$\Sigma_k = \begin{pmatrix} 1 & g_0 \\ g_0 & g_{00} \end{pmatrix}$$

$$\underline{\rho}_k' = (\rho_h, \; g_h)$$

$$M_k' = (1,1)$$

$$\underline{h} = 1$$

then

$$\tilde{\underline{a}} = \Sigma_k^{-1} [\underline{\rho}_k - M_k (M_k' \; \Sigma_k^{-1} \; M_k)^{-1} (M_k' \; \Sigma_k^{-1} \; \underline{\rho}_k - 1)]$$

$$= \frac{1}{(g_{00} - 2g_0 + 1)} \begin{pmatrix} g_{00} - g_0 - g_h + \rho_h \\ 1 - g_0 + g_h - \rho_h \end{pmatrix}.$$

Now

$$(1-\rho_h) - g_0 + g_h = (1-\rho_h) - \sum_{i=0}^{\infty} \lambda^i(\rho_i-\rho_{h+i})$$

$$+ \sum_{i=0}^{\infty} \lambda^{i+1}(\rho_i-\rho_{h+i})$$

$$= \sum_{i=0}^{\infty} \lambda^{i+1}(\rho_i-\rho_{i+1}-\rho_{h+i}+\rho_{h+i+1})$$

and since

$$g_{00} - 2g_0 + 1 = \{1 - 2(1-\lambda) \sum_{i=0}^{\infty} \lambda^i \rho_i + (1-\lambda)^2 \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \lambda^{i+j} \rho_{|i-j|}\}$$

is, apart from a multiple of σ^2, the same as

$$E\{X_t - (1-\lambda) \sum_{i=0}^{\infty} \lambda^i X_{t-i}\}^2 ,$$

i.e., the mean square error of prediction for $h = 0$, we have from theorem 4.2

$$g_{00} - 2g_0 + 1 = \frac{2\lambda}{1+\lambda} [\sum_{i=0}^{\infty} \lambda^{i+1}(\rho_i-\rho_{i+1})].$$

Hence

$$\tilde{a}_1 = \frac{(1+\lambda)}{2\lambda} (1-h(\lambda))$$

where

$$h(\lambda) = \sum_{i=0}^{\infty} \lambda^i(\rho_{h+i}-\rho_{h+i+1}) / \sum_{i=0}^{\infty} \lambda^i(\rho_i-\rho_{i+1})$$

and

$$\tilde{a}_0 = 1 - \tilde{a}_1.$$

The conditions for the mean square error to coverage to (4.46) reduce to the assumptions of a positive spectral density and the convergence of $\sum_{i=0}^{\infty} \rho_i$. The mean square error for any $\lambda \varepsilon(0,1)$ is by (4.45)

$$\sigma^2\{1 - \rho_h^2 + \frac{(\rho_h g_{00}-\rho_h g_0-g_h g_0+g_h-g_{00}-g_0^2)^2}{(g_{00}-g_0^2)(g_{00}-2g_0+1)}\}$$

and tends to

$$\sigma^2\{1-\rho_h^2\}$$

as $\lambda \to 1-$. If the process is the simple autoregressive one discussed in example 4.1 the weights turn out to be

$$\tilde{a}_0 = \{\rho^h(1+\lambda) + \lambda - 1\}/2\lambda$$

$$\tilde{a}_1 = \{(1+\lambda)(1-\rho^h)\}/2\lambda$$

compared with

$$a_0 = \rho^h$$

$$a_1 = (1-\rho^h)$$

of example 4.1.

Theorem 4.9 provides us with a method of constructing a predictor when we have a very large number of observations. Although the formulae are similar in appearance to those of theorem 3.1, the complexity of the calculations does not depends of the sample size but rather on our choice of k. The value of k we choose will depend on how many lag correlations we feel are important. If it is reasonable to assume that future observations from our process depend mostly on the immediately preceeding two or three observations we take k = 3 or 4 and the calculations will be only moderately involved. Roughly speaking, the method of prediction of theorem 4.9 is to use the most recent observations for information concerning the future values of the process and to use the remote past to estimate the unknown mean. As noted before, if the future of the process depends only on the recent past, the predictor should be quite good. At any rate, it is an improvement over the method suggested in theorem 4.8. It should be noted that the predictor in theorem 4.9 is the first one we have encountered in Chapter 4 which may be constructed to be unbiased if the mean function is a general polynomial. For small values of k and proper choices of $f(\lambda)$ the predictor will also be insensitive to long term trends in the mean of the sampled process.

Summary of Chapter 4

In this chapter we have discovered that in general there does not exist a best predictor which is unbiased and linear in the observations when we have an infinite past and an unknown mean function. We have found that the ewma of Cox depends mostly on the last ten observations in the simple autoregressive case and that the ewma may be useful in a more general setting than the simple one discussed by Cox. Bounds for the mean square error of prediction are derived for use in the general case.

Cox-type estimators of the mean are discussed and the notion of Abel-efficiency introduced. A method is given for constructing a predictor, when the mean is an unknown constant, which will be as near the optimum as desired. Formulae are derived for the best predictors in a certain restricted class of linear predictors and these predictors are shown to be arbitrarily near the optimal for a rather useful class of stationary processes.

BIBLIOGRAPHY

[1] Bailey, Martin J., "Prediction of an Autoregressive Variable Subject Both to Disturbances and to Errors of Observation," Journal of the American Statistical Association, Vol 60, (March 1965), pp. 164-181.

[2] Brown, Robert Goodell, Smoothing, Forecasting and Prediction of Discrete Time Series, Englewood Cliffs, N. J., Prentice-Hall, (1963).

[3] Cox, D. R., "Prediction by Exponentially Weighted Moving Averages and Related Methods," Journal of the Royal Statistical Society, Series B, Vol. 61, pp. 414-422.

[4] Grenander, Ulf and Rosenbaltt, Murray, Statistical Analysis of Stationary Time Series, New York, Wiley, (1957).

[5] Grenander and Szegö, Toeplitz Forms and Their Applications, Berkeley and Los Angles, University of California Press, (1958).

[6] Hajek, Jaroslav, "On Linear Estimation Theory for an Infinite Number of Observations," Theory of Probability and Its Applications, Vol. VI, No.], (1961), pp. 166-177.

[7] Hannan, Edward James, Time Series Analysis, London, Methuen, New York, Wiley, (1960).

[8] Hardy, G. H., Divergent Series, Oxford University Press, (1949).

[9] Hardy, G. H., "On the Mean Value of the Modulus of an Analytic Function," Proc. London Math Society, Sec. 2, 14, (1915) pp. 269-277.

[10] Kolmogorov, A. N., "Interpolation and Extrapolation of Stationary Random Sequences," Bulletin de l'academie des Sciences de USSR, Sec. Math. 5.

[11] Koopmans, Tjalling, ed., Statistical Inference in Dynamic Economic Models by Cowles Commission Research Staff members and guests. Introduction by Jacob Marschak, New York, Wiley, (1950).

[12] Kosulajeff, P. S., "Sur les problemes d'interpolation et d'extrapolation des suites stationnaires," Comptes Rendus de l'Academic des Sciences de l'URSS, Vol. XXX, No. 1, (1941), pp. 13-17.

[13] Quenouille, M. H., The Analysis of Multiple Time Series, New York, Hafner Publishing Company, (1957).

[14] Robinson, E. A., An Introdution to Infinitely Many Variates, Charles Griffin and Company, Ltd., London, (1959).

[15] Rosenblatt, Murray, Random Processes, New York, Oxford, University Press, (1962).

[16] Rosenblatt, Murray et al, Symposium on Time Series Analysis, Brown University, New York, Wiley, (1963).

[17] Smith, W. L., "On the Elementary Renewal Theorem for Non Identically Distributed Variables," Pacific Journal of Mathematics, 14, No. 2, (1964), pp. 673-699.

[18] Steiner, Peter Otto, An Introduction to the Analysis of Time Series, Preliminary edition, New York, Rinehart, (1956).

[19] Whittle,Peter, Hypothesis Testing in Time Series Analysis, Uppsala, Almquist and Wiksells boktr. (1951).

[20] Wiener, Norbert, Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications, [Cambridge] Technology Press of M.I.T., (1949).

[21] Wold,Herman, A Study in the Analysis of Stationary Time Series, 2nd edition, Uppsala, Almquist and Wiksells boktr., (1954).

[22] Yaglom, Akira Moiseevic, An Introduction to the Theory of Stationary Random Functions, rev. English ed., Translated and Edited by R. A. Silverman, Englewood Cliffs, N. J., Prentice Hall, (1962).

[23] Zadeh, L. A. and Ragazinni, J. R., "An Extension of Wiener's Theory of Prediction," Journal of Applied Physics, Vol. 21, (1950), pp. 645-655.