TIME-SERIES ANALYSIS: A REVIEW OF TIME-DOMAIN THEORY WITH IMPLICATIONS FOR NON-LINEAR TIME-SERIES MODELLING

by

CHRISTOPHER MAH

M.A. Carleton, 1980

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the Department

of

Mathematics

SIMON FRASER UNIVERSITY

September 1983

APPROVAL

Name: Christopher Derrick Mah

Degree: Master of Science

Title of Thesis: Time-Series Analysis: A Review of Time-Domain Theory with Implications for Non-Linear Time-Series Modelling

Examining Committee:

Chairperson: C. Villegas

Richard Lockhart
Senior Supervisor

K.L. Weldon

D.M. Eaves

T. Chang
External Examiner

Date Approved: September 16, 1983

PARTIAL COPYRIGHT LICENSE

I hereby grant to Simon Fraser University the right to lend my thesis or dissertation (the title of which is shown below) to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users. I further agree that permission for multiple copying of this thesis for scholarly purposes may be granted by me or the Dean of Graduate Studies. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Title of Thesis/Dissertation:

Time-Series Analysis: A Review of Time-Domain Theory with Implications for Non-Linear Time-Series Modelling

Author: _____
(signature)

CHRISTOPHER MAH
(name)

Sept 23 1983
(date)

Abstract

Many phenomena of interest in the sciences can be modelled by a deterministic relation perturbed by a sequence of errors in discrete time. Such a model is called a time-series. The theory of linear time-series is reviewed in detail and imperfections of well established and recent estimation methods are pointed out. It is suggested that *a priori* structural information is a neglected resource in this regard. To clarify the meaning of structural information, representation and prediction theory for weakly stationary processes is reviewed. Weakly stationary processes are seen to consist of a deterministic and a moving average part. Linear models are most suited to simulation of the moving average component. Recent proposals for non-linear time-series models are assessed in light of this discussion. It is argued that one recent proposal (threshold autoregression) is more suited to modelling of time-series with complicated deterministic parts, and requires *a priori* structural knowledge to be fully effective. In contrast another proposal (bilinear) seems more suited to time-series which are non-gaussian moving averages,and are less dependent on *a priori* structure for their usefulness.

Table of Contents

List of Tables

I. INTRODUCTION

Many phenomena of interest in the social,life and physical sciences can be usefully modelled at time-series. A time-series is a discrete-time stochastic process generated by a deterministic relation such as a difference equation, which is perturbed by errors. Theoretically, the errors are viewed either as exogenous disturbances of the deterministic dynamics of the system or alternatively, as input. Errors are modelled as random variables, which are typically assumed to be independently and identically distributed (iid).

Time-series have been of interest to statisticians because time-series data often occurs for which causal mechanisms are relatively obscure, but for which it is desirable to

i) facilitate prediction

ii) develop an insight into the causal structure of the underlying natural phenomenon

The distinction between the pursuit of (i) (predictive time-series analysis) and of (ii) (structural time-series analysis) is due to Parzen(1974). This thesis will focus on the close interrelation between these goals and we will maintain that a systematic recognition of this relationship is valuable for both purposes. This point will be illustrated for a well known special class of time-series, and later applied to the evaluation of some recent proposals for time-series models.

The time-series problem is to obtain, from a data set, a parametric form for the underlying stochastic process and estimate its coefficients. The solution of this problem demands more than the classical estimation problems of mathematical statistics because the appropriate parametric form is not specified in advance.

A time-series data set is a finite sequence of real numbers

$$\{ X(t_i) \}_{i=1}^{N} \tag{1.1}$$

obtained by observing a phenomenon of the type described above at the times t_i. When the meaning is clear from the context, a time-series data set may also be referred to as a time-series. Note that the errors occuring in the time-series model are not mentionned in this formulation of the problem, and are assumed not to be observed. In practice therefore, the errors can only be estimated from residuals, that is, from the differences between the predictions of the model and the observed values. While it can be useful to view the errors as input this is not the perspective emphasized in this thesis.

The situation in which input is observed leads to a different statistical problem from the one we consider. This alternate problem has a large literature and is appropriately treated by different methods (ie.Spectral theory) than those to be developed here. Instead, we will discuss chiefly the time-domain approach to time-series modelling. This approach deals

with a time-series directly rather than via its spectral representation.

Chapter I deals with current methods of solving the time-series problem when only linear models are admitted, and develops the theory of linear time-series models. Chapter II assesses these methods, and compares the performance of skilled versus automatic procedures on the problem of choice of autoregressive order. Because automatic methods are of special interest in the context of non-linear time-series modelling, Chapter III develops the theory of a popular automatic method, Akaike's information criterion. We conclude that automatic methods are not a panacea, and in Chapter IV demonstrate the use of structural assumptions to facilitate linear model fitting. This is, in a sense, an alternative approach to the time-series problem, and we focus on the application of this approach to non-linear modelling. Chapter V reviews representation theory for weakly stationary processes and its implication for structural modelling. Finally, Chapter VI deals with the practical application of these conclusions.

I. LINEAR MODELS:THEORY AND PRACTICE

1.1 TIME-SERIES MODELS

In attempting the time-series problem it is necessary to consider only narrow special classes of candidate models, for otherwise the problem would be intractible in practical applications. Specific hypotheses are choosen for computational convenience and so that they mirror a priori knowledge about the real processes being studied. We will enumerate the general types of times series model in common use, and discuss the solution of the time-series problem in a special case: that of the well known Box-Jenkins (linear) models. This treatment will raise general issues relevant to the structural/predictive dichotomy in non-linear time-series analysis.

Two basic types of hypothesis may be distinguished, which we will refer to loosely as models. First, models of moving average type. The most elementary form of such a model is called simply a moving average and has the form

$$X_t = q_0 \epsilon_t + q_1 \epsilon_{t-1} + \ldots + q_n \epsilon_{t-n} \qquad (1.1)$$

where t is integer, q_i are real constants, and ϵ_t is an uncorrelated mean-zero sequence of random variables.

In the notation of Box & Jenkins such a model is an MA(n)

(moving average process of order n) model and is written as

$$(1.2) \qquad X_t = \sum_{i}^{m} q_i B^i \epsilon_t \equiv Q(B)\epsilon_t$$

where the q_i are real constants, B is the backward shift operator defined by $BX_t = X_{t-1}$ and Q(B) is a polynomial defined by the second equality. If Q(B) instead of being a linear operator is some more general operator defined on the space of sequences , then the time-series is of moving average type. In considering this model it is important to recall again that the errors are not observable.

The second class of models are those of autoregressive type. In the most elementary form of such a model X_t satisfies

$$(1.3) \qquad X_t = p_1 X_{t-1} + p_2 X_{t-2} + \ldots + p_n X_{t-n} + \epsilon_t$$

where t is integer and the p_i are real constants.

In the notation of Box& Jenkins such a model is an AR(n) (autoregressive model of order n) process and it is often written

$$(1.4) \qquad \sum_{i}^{n} p_i B^i X_t \equiv P(B)X_t = \epsilon_t$$

where P(B) is a polynomial defined by the first equality. We will refer to the term

$$B^m X_t = X_{t-m} \tag{1.5}$$

as the term corresponding to the lag of order m.

If P(B) instead of being a linear operator is some more general operator defined on the space of sequences , then the time-series is of autoregressive type. Time-series which satisfy relations of the form

$$P(B)X_t = Q(B)\epsilon_t \tag{1.6}$$

where the symbols are as above, are said to be of mixed type.

The elementary linear versions of these models were introduced by Yule(1927) and are supported by a large body of theory (Box&Jenkins,1970:Koopmans,1974; Deutsh,1965), as well as a set of relatively refined estimation techniques. For example, many commonly available time-series analysis programmes (such as the IMSL routine FTCMP) are devoted to the estimation of parameters in Box-Jenkins models.

1.2 BOX-JENKINS MODELS: INVERTIBILITY AND STATIONARITY

We now begin a discussion of the solution of the time-series problem in the linear case. To do this it is necessary to describe the theory of Box-Jenkins autoregressive/moving average (ARMA) models. We begin with the definition of the natural sample statistics to compute if one is interested in the linear predictibility of the observations. Consideration of these definitions leads to assumptions on the stochastic processes so that such procedures make sense; namely weak

stationarity and invertibility. The meaning and mathematical interpretation of these conditions is discussed. Finally, it will be shown that under these assumptions the covariance structure of a process leads to a nearly unique determination of the coefficients of an ARMA model.

In examining an empirical data series and assessing its predictibility it is natural to compute the autocorrelation function of the observations. This is defined by

$$(2.1) \qquad r_j = \frac{N}{N-j} \sum_{i=1}^{N-j} (x_{i+j} - \bar{x})(x_i - \bar{x}) \Big/ \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})$$

where N is the length of the series and $\bar{x}$ is the series mean.

This statistic summarizes the extent to which we can (within the sample) linearly predict X_{t+j} from X_t. For this procedure to give a result generalizable to another sample from the same time-series, it must be assumed that

$$(2.2) \qquad E[X_t] \equiv \mu = \int X_t(\omega) dP = \lim_{T\to\infty} \frac{1}{2T} \sum_{-T}^{T} X_t < \infty$$

independent of t and

$$(2.3) \qquad \rho_j = \int (X_{t+j}(\omega) - \mu)(X_t(\omega) - \mu)\, dP$$

$$= \lim_{T\to\infty} \frac{1}{2T} \sum_{T}^{-T} (X_{t+j} - \mu)(X_t - \mu) < \infty$$

The first equalities for all t comprise the assumption of weak stationarity while the second set of equalities (limits being taken in the L^2 sense) characterize ergodicity. The assumption of weak stationarity may be contrasted with (strong) stationarity, for which the joint probability distribution of any finite collection rather than only the first two moments, must be independent of t. Thus the process X_t is strongly stationary if and only if the joint distribution function of the collection

$$\{ X_{t+s_1}, X_{t+s_2}, \ldots X_{t+s_n} \} \tag{2.4}$$

does not depend on t.

Much of the treatment of Box-Jenkins models assumes only weak stationarity of the error series. In particular, it is assumed that $E[e_t] = 0$ and $E[e_t e_{t-k}]$ is constant in t for each k, and zero for k=0. In general this implies only weak stationarity of the resulting series X_t. However, should be noted that when the error series is normally distributed, a Box-Jenkins series X_t has multivariate normal joint distributions because of linearity. Thus, in this case weak stationarity of the errors implies strong stationarity of the model.

A moving average is always stationary. If the errors become independent sufficiently rapidly, it is also ergodic. Suppose for example that ϵ_t is independent of ϵ_{t-k} for k>K , all joint moments of the sequence exist, and the variance is constant. Now if K=1 (the sequence is independent) then the ϵ_t sequence is mean-ergodic because, the variance of the sum $\Sigma_t \epsilon_t$ is $N\sigma_\epsilon^2$. Thus the variance of the average is order $\frac{1}{N}$ and so, tends to zero as N-->∞ . Suppose now K > 1. Computing the variance of the sum using the assumptions we find that the number of non-zero covariances in the computation grows as NK. Thus, since each covariance is bounded by σ_ϵ^2 the variance of the average is again order $\frac{1}{N}$ and so tends to zero. Covariance ergodicity is obtained similarly on computing the variance of the sum $\Sigma_t \epsilon_t \epsilon_{t-1}$. This shows that the error sequence is ergodic in the sense of (2.2-3), and from this it follows that finite moving averages are ergodic. A further easy argument extends this result to infinite moving averages which converge geometrically, as the variance of the tail of the sum tends to zero.

It is a comparatively simple matter to compute the theoretical autocovariance function γ_j for an MA(n) process. It is computed as

$$\gamma_0 = \sigma_\epsilon^2 \sum_i q_i^2 \tag{2.5}$$

$$\gamma_m = \sigma_\epsilon^2 \sum_i q_i q_{i+m} \qquad \text{for } m \leq n$$

$$\gamma_m = 0 \qquad \text{for } m > n$$

where σ_ϵ^2 is the variance of the ϵ_t.

It is a little more difficult to compute the autocovariance function of a (stationary) autogressive process and this will be postponed until later. For now we wish to deduce its behavior for large lags in order to contrast it with an moving average process. To do this notice that because X_t satisfies the difference equation (1.4), the autocovariances $\text{Cov}(X_t, X_{t-n}) = \gamma_{t,n}$ satisfy the difference equation

$$\gamma_{t,m} - q_1 \gamma_{t,m-1} + \ldots + q_m \gamma_{t,m-n} = 0 \tag{2.6}$$

To see this multiply the original equation by X_{t-n} and take expectations. Because of stationarity, the t subscript may be omitted and it is seen that the autocovariance sequence $\{\gamma_m\}$ satisfies the homogeneous difference equation. The difference

equation (2.6) can be solved to yield a sequence which is a linear combination of sequences of the form

(2.7) $$\gamma^t \quad \text{or} \quad t^k \gamma^t$$

For notice that the operator P(B) may be resolved into

(2.8) $$(I - r_1 B)(I - r_2 B) \ldots (I - r_n B)$$

where I is the identity operator and B is the backward shift operator, on using the fundamental theorem of algebra, and the resulting factors commute. Then substitute terms of the form above to see that they are annihilated by the resulting operator. The conclusion is that autocorrelations of an AR process must grow or decay exponentially with time. This is explained more fully in appendix A.

1.2.1 STATIONARITY

We have seen that a moving average process is automatically stationary. In contrast there are non-trivial conditions for an autoregressive process to be stationary. For consider the time-series

(2.9) $$X_t = pX_{t-1} + \epsilon_t$$

where $p \neq 1$, the ϵ_t are a stationary zero mean process and ϵ_t is independent of X_{t-1}. The variance of X_t can be found by squaring both sides of this equation and taking expectations. Doing this, bearing in mind the independence of X_{t-1} and ϵ_t we find

$$E[X_t^2] = p^2 E[X_{t-1}^2] + \sigma_\epsilon^2 \tag{2.10}$$

where σ_ϵ^2 is the variance of the ϵ_t. If $\{X_t\}$ were stationary $E[X_t^2] = \gamma_o$ a constant, and the equation becomes

$$\gamma_o = p^2 \gamma_o + \sigma_\epsilon^2 \tag{2.11}$$

Inspecting equation (2.11) it is seen that there is no constant positive solution for γ_o when $|p|>1$. Thus no process satisfying the difference equation (2.9) with $|p| > 1$ and ϵ_t independent of X_{t-1} is weakly stationary.

If $|p|<1$, then the sum $\sum_{k=0}^{\infty} p^k \epsilon_{t-k}$ defines a stationary ergodic random process with a finite variance. It is easy to see that this satisfies equation (2.9). Is this solution unique? To answer this question note that $E[X_t^2] = \gamma_o$ is uniquely determined by (2.11). Moreover,because the covariances satisfy the homogeneous difference equation

$$\gamma_{t+1} = p\gamma_t \tag{2.12}$$

whenever $\mathrm{Var}[X_t] = E[X_t^2]$ is known, $\mathrm{Cov}[X_t, X_{t-k}]$ for $k>0$ are uniquely defined. Thus, there is a weakly stationary process

with unique covariance structure satisfying the difference equation with $|p|<1$.

When the errors are iid this is the stochastic process which results from starting the Markov chain X_t in its stationary distribution. Henceforth, this situation will be summarized as; an AR(1) process is stationary iff $|p|<1$. Note that we now say stationary, and understand, weakly stationary.

Now this is not the whole story on stationarity of linear processes, for if the above procedure is repeated under the assumption that ϵ_t is independent of X_t in the defining difference equation, it is then found that the resulting process is stationary iff $|p|>1$. This point will be clarified in section III but first we will extend the above reasoning to more complicated AR processes.

Observe that if X_t is a stationary process so is $P(B)X_t$ for any P. To see this just write out the operator, multiply by any time translate and take expectations. This fact, together with the discussion of the above process implies that

(2.13) $$P(B)X_t = e_t$$

can define a weakly stationary process only when the roots of $P(x)=0$ lie outside the unit circle (note that the roots of $P(x)=0$ are the reciprocals of the numbers r_i in the resolution of the operator $P(B)$). For resolve $P(B)$ as

(2.14) $$P(B) = (I-r_1 B)(I-r_2 B)\ldots(I-r_n B)$$

We procede by contradiction. Suppose that some $|r_i| \geq 1$, say for $i=1$ and that X_t is stationary. Define

$$T(x) = (I - r_1 x)^{-1} P(x) \tag{2.15}$$

and set

$$Z_t = T(B)X_t \tag{2.14}$$

Then Z_t satisfies

$$\epsilon_t = (I - r_1 B)Z_t \tag{2.16}$$

so by the discussion Z_t is non-stationary, but being a 'polynomial' in lags of X_t it is stationary by the hypothesis. This gives a contradiction, and the assertion is proved. The implications of roots on the unit circle will be discussed in a later section dealing with non-stationary behavior.

Finally note that if $|r_i| < 1$ for $i=1,n$ then the operator $P(B)$ may be inverted, factor by factor using the formula

$$(I - r B)^{-1} = I + rB + r^2 B^2 + \ldots \tag{2.17}$$

to obtain an infinite moving average representation for the series which converges in probability and solves the difference equation (2.13) . In section 1.3 it will be seen that the covariance structure of any solution must be the same as the covariance structure of this sum.

1.2.2 INVERTIBILITY

We are now prepared to consider the notion of invertibility of a process. The invertibility of a process is an algebraically dual condition to stationarity which also has a practical meaning. Consider the mixed process

$$P(B)X_t = Q(B)e_t \tag{2.18}$$

As shown above, the condition for stationarity is that $P(x)=0$ have roots lying outside the unit circle. The algebraic condition for invertibility of a process is that $Q(x)=0$ have roots lying outside the unit circle. One consequence of this will clearly be that $Q(B)$ may be inverted to yield a purely autoregressive representation

$$e_t = Q^{-1}(B)\, P(B)X_t \tag{2.19}$$

for the mixed process.

The invertibility condition may also be interpreted as follows: Suppose that $X(t)$ is a series of interest and that at time $t=n$ a forecast $f(n,h)$ is required of the future value $X(n+h)$. If a least square criterion is adopted then the optimal forecast is

$$f(n,h) = E\{ X(n+h) \mid \text{past}\} \tag{2.20}$$

or in other words

$$f(n,h) = \sum_i p_i f^{x}(n,h-i) + \sum_j q_j f^{e}(n,h-j) \tag{2.21}$$

where

$$f^{x}(n,h-j) = \begin{cases} X(n+h-j) & \text{for } j \geq h \\ f(n,h-j) & \text{for } j < h \end{cases}$$

and

$$f^{e}(n,h-j) = \begin{cases} 0 & \text{for } j < h \\ e(n+h-j) & \text{for } j \geq h \end{cases}$$

Thus, consider the MA process

(2.22) $$X_t = e_t + qe_{t-1} = (I + qB)\, e_t$$

where e_t is a stationary zero mean process. On applying the above principle conclude

(2.23) $$f(n,1) = q\, e_n$$

however, the e_t were not observed and we know only X_t. Thus e_t must be estimated from previous values of X_t.

We now investigate the feasibility of this as the data series becomes longer. First note that if there was a good guess (or any guess)

$$\hat{e}_0 \text{ for } e_0$$

available then the recursion could be used to get a sequence of estimates

(2.24) $$\hat{e}_t = X_t - q\, \hat{e}_{t-1}$$

since X(t), t=1,N were observed. It turns out that this is

only sensible when $|q|<1$. To see this define the inaccuracy h_t as

$$h_t = \hat{e}_t - e_t \tag{2.25}$$

then

$$h_t = \{ [X_t] - q\hat{e}_{t-1} \} - e_t$$

$$= [e_t + qe_{t-1}] - q\hat{e}_{t-1} - e_t = - q h_{t-1}$$

Thus, using this method, on a long data series, the inaccuracy in an initial guess dies out when $|q|<1$ and fails to die out when $|q| \geq 1$. Because the errors are not observed, this is the only possible method of forecasting. Therefore the practical interpretation of invertibility is that the optimal forecast may be approached arbitrarily closely with sufficiently long data series. The above example (2.22) is invertible iff $|q|<1$.

For a more complicated MA process

$$X_t = Q(B)\, e_t \tag{2.26}$$

The optimal forecast (as noted) still involves e_t. But for any initial guess $\hat{e}_0$ we have

$$Q(B)\, h_t = 0 \tag{2.27}$$

and the solutions of this difference equation are stable only if the roots of $Q(x)=0$ lie outside the unit circle. When this

condition is satisfied it is again true that the inaccuracies h_t in the choice of $\hat{e}_t$ die out, and the optimal forecast may be approached arbitrarily closely.

1.3 BOX-JENKINS MODELS:COVARIANCE STRUCTURE AND UNIQUENESS OF THE MODEL

To discuss covariance structure it is most convenient to have the model in moving average form. We claim that any mixed ARMA model $P(B)X_t = Q(B)e_t$ for which $|r_i| \neq 1$ in the resolution of P, can be written

$$X_t = \sum_{-\infty}^{\infty} v_j e_{t-j} \equiv V(B)\, e_t \tag{3.1}$$

where V(x) is a formal power series. For certainly the factors $I - rB$ in the resolution of P with $|r| < 1$ may be inverted as noted above, while for each factor with $|r| > 1$ rewrite as

$$(I - rB) = BFr\,(I\,r^{-1} - B) \tag{3.2}$$

$$= -B\,r\,(I - r^{-1}F)$$

where B is the backward shift operator and F is the forward shift operator, so that

$$(I - rB)^{-1} = -(I - r^{-1}F)^{-1}\, r^{-1}\, F^{-1} \tag{3.3}$$

In this manner it becomes meaningful to write

$$X_t = P^{-1}(B)\, Q(B)\, e_t = V(B)\, e_t \tag{3.4}$$

The autocovariances in this model can be computed as

$$(3.5) \qquad E[X_{t+j} X_t] = \sum_k \sum_l v_k v_l E[e_{t+j-k} e_{t-l}]$$

$$= \sigma_\epsilon^2 \sum_l v_l v_{l+j} = \gamma_j$$

Next, define the autocovariance generating function

$$(3.6) \qquad \Lambda(B) = \sum_j \gamma_j B^j$$

Substituting the previous expression for the autocovariances obtain

$$(3.7) \qquad \Lambda(B) = \sigma_\epsilon^2 \sum_j \sum_l v_l v_{l+j} B^l = \sigma_\epsilon^2 \sum_j \sum_k v_j v_k B^{j-k}$$

$$= \sigma_\epsilon^2 V(B)V(F)$$

We can now show that for a given autocovariance generating function, the resolutions of the operators P(B) and Q(B) are defined up to reciprocals of the roots. For the autocovariance generating function corresponding to (3.4) is

$$(3.8) \qquad \sigma_\epsilon^2 \{ P^{-1}(B) Q(B) \} \{P^{-1}(F) Q(F) \}$$

using (3.7).

Thus, from the two operator identities(3.9)

$$(i) \quad (1 - \phi_i B)(1 - \phi_i F) = \phi_i^2 (1 - \phi_i^{-1} B)(1 - \phi_i^{-1} F)$$

$$\text{(ii)}\quad (1 - \psi_i B)^{-1} (1 - \psi_i F)^{-1} = \psi_i^{-2} (1 - \psi_i^{-1} B)^{-1} (1 - \psi_i^{-1} F)^{-1}$$

where ψ_i and ϕ_i are real constants, it is seen that all the processes

$$\text{(3.10)}\qquad \prod_{i=1}^{m} (1 - \psi_i^{\pm 1} B) X_t = \delta \prod_{i=1}^{n} (1 - \phi_i^{\pm 1} B) e_t$$

where δ is chosen to scale the variance appropriately have the same autocovariance generating function and hence the same autocovariances.

Thus, the restrictions of invertibility and stationarity allow the choice of exactly one of the reciprocal pairs of each factor (provided the roots are off the unit circle) and ensure that the model corresponding to a given covariance structure is unique up to cancellation of factors in the equation.

It is important to note that the non-uniqueness discussed above is not a non-uniqueness in distribution. In fact when the errors are normally distributed, knowledge of the covariance function is equivalent to knowledge of the joint distributions of

$$\text{(3.11)}\qquad (X_{t_1}, X_{t_i}, \ldots, X_{t_n})$$

where t_i is any finite collection of times. Rather, the non-uniqueness arises from the fact that the Box-Jenkins representations contain information about the relation of past,

present and future. No restrictions follow from this when the defining relationship involves a finite number of random variables. For, if the present is time t=n then a stationary process which satisfies

$$X_{n+m} = p_o X_{n+m-1} + p_1 X_{n+m-2} + \ldots + p_n X_n \tag{3.12}$$

for each n, must also satisfy

$$X_n = p_o X_{n-1} + p_1 X_{n-2} + \ldots + p_m X_{n-m} \tag{3.13}$$

But the situation is different for one-sidedly infinite relations. Consider the stationary AR(1) process

$$X_t = qX_{t-1} + e_t \tag{3.14}$$

Recall that this can be written as

$$(I - qB) X_t = e_t \tag{3.15}$$

and this implies (for $|q| \leq 1$)

$$X_t = e_t + qe_{t-1} + q^2 e_{t-2} + \ldots \tag{3.16}$$

Hence, by a straightforward calculation we see that when the autocovariance generating function makes sense (ie. when X_t is stationary) it depends only on q. There is no non-uniqueness here. However it was seen above that for each value of q there is a Box-Jenkins time-series equation giving the same autocovariance generating function which depends on 1/q. Does

this equation generate a distinct stationary process? Recall that in the discussion of stationarity conditions it was stated that if in an AR(1) difference equation e_t is assumed independent of X_t, the resulting process is stationary iff $|q|>1$. Suppose then that we assume $|q|>1$ and e_t independent of X_t. Rewrite the difference equation for this stationary process as

$$(I - F/q)\, X_t = w_t \tag{3.17}$$

where F is the forward shift operator and

$$w_t = -e_t / q \tag{3.18}$$

then since $|q|>1$ this process may be inverted as a function of future errors:

$$X_t = w_t + q^{-1} w_{t+1} + q^{-2} w_{t+2} + \ldots \tag{3.19}$$

Thus there are two distinct processes with the same autocorrelation function . While the two processes above will have the same autocovariances and possibly the same joint distributions, one describes a process independent from past e_t's and the other a process independent of future e_t's. Fortunately the distinction is merely philosophical since the errors are not observed, so that there is no way to distinguish these models from data. It makes good sense however, to choose

the representation in which a convergent expansion in terms of the past is stationary.

In a parallel fashion consider the MA(1) process

$$X_t = (I - pB)\, e_t \tag{3.20}$$

this may be inverted as (for $|p|<1$)

$$e_t = X_t + pX_{t-1} + p^2 X_{t-2} + \ldots \tag{3.21}$$

There is another Box-Jenkins model which gives the same autocovariance generating function, which is constructed by replacing B with F and p by 1/p and multiplying by an appropriate scale factor. To verify that a distinct stationary process is generated note that if $|p|>1$, the original equation may be rewritten as

$$Z_t = (I - p^{-1} F)\, e_{t-1} \tag{3.22}$$

where F is the forward shift operator and $Z_t = -X_t / p$

and may be 'inverted' as

$$e_{t-1} = Z_t + p^{-1} Z_{t+1} + p^{-2} Z_{t+2} + \ldots \tag{3.23}$$

Thus in the first case (AR(1)) the resulting process X_t depends on its past and a current error e_t while in the second (MA(1)) X_t depends on its future and a current error e_t. This

ambiguity is again resolved by appealing to physical intuition.

1.4 NON-STATIONARY MODELS

Box-Jenkins models provide naturally for non-stationary behavior. The most useful type of non-stationarity occurs when $|r| = 1$ in one or more of the factors $I-rB$ in the polynomials P,Q. First note that if $r = 1$ that the factor $I-rB$ becomes the difference operator D defined by

$$DX_t = X_t - X_{t-1} \tag{4.1}$$

This suggests a natural model for data which locally exhibits a stochastic polynomial trend over time. Differencing once transforms a sequence which grows linearly to a sequence of constants, and differencing twice takes a sequence growing quadratically to a sequence of constants. Therefore sequences which contain polynomial trends of degree d but which become stationary when the trends are removed, may be modelled as the solutions of

$$P(B)\, D^d X_t = Q(B)\, e_t \tag{4.2}$$

If $d=1$ this model is a generalization of a random walk model in which the increments, instead of being independent, are a Box-Jenkins process.These models $d\geq 1$ are known as integrated autoregressive moving average models and are identified by the notational convention ARIMA(p,d,q) where p and q are respectively the autoregressive and moving average orders and d

is the order of differencing. On writing down the associated difference equations for the autocovariances as we did in section 1.2, it is seen that the autocorrelation function for an ARIMA model with $d \geq 1$ will tend to grow rather than die out, and it is the failure of the sample autocorrelation function to die out that is taken as an indication of non-stationarity in practice. On the other hand if $r = \cos[\theta] \pm i \sin[\theta]$, $r \neq 1$ then this implies periodic behavior of the series, for example if $|r_i| = 1$, $r_i \neq 1$, $i=1,2$ and

$$P(B)X_t \equiv (I - r_1 B)(I - r_2 B) X_t = e_t \tag{4.3}$$

then the null space of the difference operator includes functions such as $\cos[\theta t]$ and $\sin[\theta t]$, as the real and imaginary parts of r^t must separately solve the homogeneous equation. This means the difference equation admits non-stationary solutions such as

$$\cos[\theta t] + w_t \tag{4.4}$$

where w_t is any solution of the defining relation (4.3).

In this manner it is possible to model data which contains periodic components with random changes in amplitude and phase. Because the autocovariances must satisfy the homogeneous difference equation, it may be seen as before that the autocorrelation function of such a process will fail to die out.

1.5 SOLVING THE TIME SERIES PROBLEM

Box&Jenkins (1970) distinguish the following stages of model building:

(1) Identification, in which we attempt to determine a class of models which may be sensibly entertained in the light of a priori considerations and a rough data analysis, but is small enough that reasonably efficient parameter estimates are available

(2) Estimation, in which, having decided upon a class of models we form estimates of the parameters of the models which is supposedly being observed

(3) Diagnostic checking, in which by residual analysis it is tested whether the model is consistent with the observations post hoc.

We now discuss each of these stages in turn, sketching the general procedures involved. The aim of this section is to demonstrate that the linear time-series problem is a difficult statistical problem which leads to many unresolved issues. It will become clear that the fitting of a Box-Jenkins model, even when the basic assumptions are correct may leave much room (relative to other statistical procedures) for judgement. It will be suggested that a somewhat neglected method of coping with such difficulties is to make more deliberate use of structural information. A discussion of some structural properties of Box-Jenkins models follows which will later be contrasted with the properties of non-linear time-series models.

1.5.1 IDENTIFICATION

When it is known in advance what the correct form of the linear model is, estimating the coefficients of a time-series is relatively straightforward. Unfortunately such knowledge is rarely available in convenient form. Thus the goal in the early stages of data analysis is to decide whether a time-series is purely a moving average, purely autoregressive or mixed and to determine the orders of the components. To decide initially on the correct class of linear models it is usual to begin by computing the autocorrelation function and the partial autocorrelation function of the observations.

The partial autocorrelation function is defined as follows. Let $H_{t,s}$ denote the set of random variables $\{X_{s+1},\ldots,X_{t-1}\}$ for $t > s+1$ and let $\hat{X}_s$ and $\hat{X}_t$ denote respectively the predictions of X_s and X_t from the true regressions of X_s and X_t on $H_{t,s}$. Then the $t-s^{th}$ partial autocorrelation of a stationary process is defined as

$$(5.1) \qquad \phi_{t-s} = \text{Correlation}[X_t - \hat{X}_t\ ,\ X_s - \hat{X}_s]$$

where we agree that when both the arguments of the correlation are constant we will set $\phi_{t-s} = 1$. When both the arguments of the latter correlation are non-constant random variables the n^{th}

partial autocorrelation is

(5.2) $\hat{p}_n^n$ = the last coefficient in the true multiple regression of X_t on $X_{t-1}, X_{t-2}, \ldots X_{t-n}$

We have seen that the theoretical autocorrelations for an MA process have a cutoff, while the autocorrelations of an AR process are non-zero for large lags. Now the latter statement holds with autocorrelations replaced by partial autocorrelations and AR and MA interchanged. Thus, the computation of the autocorrelation function and partial autocorrelation function is a preliminary procedure for distinguishing MA and AR processes. The claim is proved as follows. Consider the autoregression

$$e_t = \sum_{i=0}^{n} p_i X_{t-i} \tag{5.3}$$

where $p_o = 1$, the series converges in probability and possibly $n=\infty$. If the object is to minimize

$$E[(X_t - \hat{X}_t)^2] \tag{5.4}$$

where $\hat{X}_t = \hat{p}_1 X_{t-1} + \hat{p}_2 X_{t-2} + \ldots + \hat{p}_m X_{t-m}$ for $m \geq n$ then we must minimize

$$\text{Var}\{ (\hat{p}_1 - p_1) X_{t-1} + (\hat{p}_2 - p_2) X_{t-2} + \ldots + (\hat{p}_n - p_n) X_{t-n} + \hat{p}_{n+1} X_{t-n-1} + \ldots + \hat{p}_m X_{t-m} + e_t \} \tag{5.5}$$

Since the X_t are jointly non-degenerate random variables, and e_t is uncorrelated with X_{t-k} for $k>0$ conclude

$$\hat{p}_i = p_i \quad \text{for } i=1,m \tag{5.6}$$

and in particular $\hat{p}_i = 0$ for $i > n$. Thus for $i > n$, where n is the order of the autoregressive process $\phi_i = 0$.

When $m < n$ the latter variance (5.5) becomes

$$\begin{aligned} \text{Var}\{ &(\hat{p}_1 - p_1) X_{t-1} + (\hat{p}_2 - p_2) X_{t-2} + \ldots \\ &+ (\hat{p}_m - p_m) X_{t-n} + p_{m+1} X_{t-m-1} + \ldots + p_n X_{t-n} + e_t \} \end{aligned} \tag{5.7}$$

and the first argument fails. In fact, some $\hat{p}_i \neq p_i$ unless $p_{m+1} X_{t-m-1} + \ldots + p_n X_{n\,t} + e_t$ is uncorrelated with $X_{t-1}, \ldots, X_{t-m}$ but these quantities are known to be correlated. Therefore we note that the 'true' coefficients of the multiple regression of X_t on $X_{t-1}, \ldots, X_{t-m}$ for $m < n$, are not the coefficients of the underlying Box-Jenkins process.

The preceding argument shows that when a process is autoregressive of finite order, the partial autocorrelation function cuts off, and when the process is a moving average, n is infinite and so the partial autocorrelation function is generally non-zero for arbitrarily large lags as claimed. Ramsey(1974) has characterized the partial autocorrelation

function of a stationary process by the following striking result. Define the set of sequences S as

$$(5.8) \quad S = \{[\phi_i] \in R^{\infty} : |\phi_i| \leq 1 \text{ and } |\phi_i| = 1 \Rightarrow \phi_{i+1} = 1 \}$$

then a function ϕ_i belongs to S iff it is the partial autocorrelation function of a stationary process. A time-series whose partial autocorrelation function is eventually unity belongs to the class of singular processes, which are processes perfectly predictable in mean square.

Because the partial autocorrelation function arises in the section on AR order selection some remarks on its computation are in order. This discussion will also gives some additional insight into the nature of the partial autocorrelation function. For this purpose we introduce the Yule-Walker equations.

When the process concerned is purely a moving average or purely autoregressive there is a simple relationship between the autocovariances and the parameters of the Box-Jenkins process. If the process is autoregressive then the parameters may be obtained uniquely from the Yule-Walker equations which are obtained as follows. Suppose the process is of order m. Then multiplying both sides of the defining difference equation by X_{t-k} for k=1,m and taking expectations obtain

$$\gamma_1 = p_1\gamma_0 + p_2\gamma_1 + \ldots + p_m\gamma_{m-1} \tag{5.9}$$

$$\vdots$$

$$\gamma_i = p_1\gamma_{i-1} + p_2\gamma_{i-2} + \ldots + p_m\gamma_{m-i}$$

$$\vdots$$

$$\gamma_m = p_1\gamma_{m-1} + p_2\gamma_{m-2} + \ldots + p_m\gamma_0$$

which may be solved directly for p_i $\quad i=1,m$

This leads to a natural estimate for the p's (provided that the order of the autoregression is known) obtained by replacing the theoretical autocorrelations above by their sample estimates. This estimate (called the method of moments estimate) is used in practice as an initial guess for maximum likelihood computations (cf. Box& Jenkins,1970; Appendix 6.2).

In a similar fashion, it was shown in section 1.2 that for a moving average process of order m the autocovariances are

$$\gamma_k = (q_1 q_{1+k} + q_2 q_{2+k} + \ldots + q_{m-k} q_m)\,\sigma_e^2 \tag{5.10}$$

where σ_e^2 is the variance of the errors. These non-linear equations may be solved for q_i $\quad i=1,m$. While the non-linear equations will have multiple solutions it follows from the discussion of section 1.3 that there is a unique invertible Box Jenkins model corresponding to a given autocorrelation function.

We now return to the partial autocorrelation function . It is not hard to see that if the Yule-Walker system has an order less than the true one, then the resulting coefficients are the coefficients $\hat{p}_i^m$ in the multiple regression of X_t on $X_{t-1}, \ldots, X_{t-m}$. This is because to minimize the variance $Var[X_t - \hat{X}_t]$ we compute this variance as a function of the p's (a quadratic form with coefficients γ_i and variates $\hat{p}_i^m$) and set its gradient with respect to the p's to zero. This results in a set of formulas for the partial autocorrelation function in terms of determinants. The linearity of this problem enables an easy computation of the residual variance .

For let $H_n = \text{span}\{X_{t-1}, \ldots, X_{t-n}\}$, $x = \hat{X}_t$ and $y = X_t$. Define the inner product $[z,v] = E\{ zv \}$. With these definitions, the problem is equivalent to the Hilbert space problem of finding $S_n = \underset{x \epsilon X}{\text{Min}} \| y - x \|$ where X is a subspace of H_n of dimension m.

It is well known that the solution x^* to this problem is characterized by $[x^* - y, x^*] = 0$ and that the norm of x^* is the norm of the projection of y on H_m . The latter projection is equal to

$$(5.11) \qquad E[X_t (p_1 X_{t-1}, \ldots, p_m X_{t-m})] = \sum_{k=1}^{m} p_k \gamma_k$$

where p_k are the true coefficients of the autoregressive representation of the process. Thus, applying the pythagorean theorem we obtain

$$(5.12) \qquad S_m = \gamma_o - \sum_{k=1}^{m} p_k \gamma_k .$$

as the residual of the autoregression of order n.

We can now state the recursive formulae due to Durbin(1960) for computation of the partial autocorrelation function . As noted earlier, $\hat{p}_m^m$ is the m^{th} partial autocorrelation of X_t. For ease of notation we now drop the ^ on $\hat{p}_i^m$, since the distinction is clear from the superscript. Durbin's algorithm is a recursive joint computation S_m and p_k^m for k=1,m . The recursion begins

$$(5.13) \qquad S_o = \gamma_o$$

$$p_1^1 = \frac{\gamma_1}{\gamma_o}$$

$$S_1 = \{ 1 - (p_1^1)^2 \} S_0$$

and the general steps are

$$(5.14) \qquad p_n^n = \{ \gamma_n - \sum_{k=1}^{n-1} p_n^{n-1} \gamma_{n-k} \} / S_{n-1}$$

$$p_k^n = p_k^{n-1} - p_n^n p_{n-k}^{n-1} \qquad \text{for} \qquad k=1,n-1$$

$$S_n = \{ 1 - (p_n^n)^2 \} S_{n-1}$$

From the last equation the necessity of the conditions on the set S above are evident. For clearly $|p_n^n| \leq 1$ and if $|p_n^n| = 1$ this implies $S_{n+1} = 0$. When $S_{n+1} = 0$ this means that the process satisfies a homogeneous difference equation of order n+1 or less. Thus the correlation in the definition of ϕ_{n+1} reduces to a correlation between identically zero variates and so, by definition, is unity. Ramsey's result arises by observing that these equations can be used to generate an admissible sequence of autocovariances from the partial autocorrelation function .

His result is not completely trivial, for in contrast to the weak conditions on a sequence of numbers which make it admissible as a partial autocorrelation function , many sequences cannot be the autocorrelation function of a stationary process. Thus, for example, there are non-vaccuous conditions on a set of m autocorrelations generated by a moving average process in addition to exhibiting cut off. For example given an MA(1) process

(5.15) $$X_t = e_t + q e_{t-1}$$

compute

$$\gamma_0 = E[X_t X_t] = \{1 + q^2\}\sigma_e^2 \tag{5.16}$$

$$\gamma_1 = E[X_{t-1} X_t] = E\{e_{t-1} + qe_{t-2}\}\{e_t + qe_{t-1}\}$$

$$= q\sigma_e^2$$

so that

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{q}{(1+q^2)} \tag{5.17}$$

and solving for q in terms of ρ_1 we find q is real iff $|\rho_1| \leq .5$. Since $|\rho_1| = .5$ implies $q = 1$, the resulting process is an invertible moving average iff the inequality is strict.

A general necessary and sufficient condition for a sequence of numbers ρ_i , i=1,..m to be the autocorrelation function of an invertible moving average process was given by Wold (1953). Define

$$U(x) = 1 + \sum_{j=1}^{m} \rho_j (x^j + x^{-j}) \tag{5.18}$$

now define

$$V(z) = U(x) \quad \text{if} \quad z = x + x^{-1} \tag{5.19}$$

Then the condition is: V(z) has no root in the interval [-2,2].

We sketch the proof. To begin, we recall that from the discussion of section 1.2, any m^{th} degree polynomial Θ

corresponds to some moving average model $X[\Theta]$. Such a moving average model is always stationary, but conceivably not invertible. For invertibility of the associated moving average model, the roots of the operator resolution of Θ must lie in the interior of the unit disk of the complex plane. However, the discussion of section 1.3 shows that if one or more roots lie in the interior of the complement of the disk, a new (unique) invertible moving average process may be defined with r replaced by $\frac{1}{r}$ possessing the same autocovariances. Thus any choice of Θ such that the roots lie off the unit circle corresponds to a unique invertible moving average (and thus a set of m moving average autocorrelations). On the other hand it can be shown that any sequence of m autocorrelations corresponds to an m^{th} degree Box-Jenkins (moving average) operator Θ. The preceding remarks thus show that a given autocorrelation sequence corresponds to an invertible moving average process iff roots of the associated operator Θ do not lie on the unit circle. Our object is to translate this condition on Θ into a condition on the autocorrelation function (or equivalently, the autocovariance function) of the process.

To do this first recall that the autocovariances of a moving average process are $\gamma_t = \sigma^2 \Sigma_j q_j q_{t+j}$. Inspection of this formula shows that we can represent the autocovariance sequence alternatively as follows. Define the fourier transform of a

sequence q_i, $i=0,m-1$

$$\Theta(\phi) = \sum_{j=0}^{m-1} q_j e^{ij\phi} \tag{5.20}$$

Set $x = e^{i\phi}$ and consider (abusing notation slightly)

$$\Gamma(\phi) = \Theta(x) \cdot \Theta^{*}(x) \tag{5.21}$$

where the * denotes complex conjugation. The result of this computation is certainly real and can be shown to equal twice the cosine transform of the autocovariance sequence. Thus the autocovariances could be recovered from Γ. This representation is useful because on inspecting (5.21) we notice $\Gamma(\phi) = 0$ iff $\Theta(x) = 0$. Thus, as $\Theta(x)$ is just the Box-Jenkins operator evaluated on the unit circle, we can detect the presence of roots of Θ on the unit circle by inspecting Γ, which is twice the cosine transform of the autocorrelation sequence of length m. Therefore our aim of translating the condition on the roots of Θ has been acheived. To complete the proof, note that the transformation $z = x + x^{-1}$ maps x on the unit circle to the real interval [-2,2].

For example, in the model (5.15) we obtain for U(x), on substituting q for q_1

$$[(1 + q^2) + 2q\cos(\phi)] = [1 + \rho_1 (x + x^{-1})] \tag{5.22}$$

In this example it is clear that that to avoid a root on the unit circle in (5.22) $|\rho_1| < .5$. This is the same condition obtained earlier in this section.

1.5.2 MODEL ESTIMATION

Once the order of the ARMA processes have been specified it becomes feasible to estimate the parameters of the model. The usual method is the method of maximum likelihood or some variation of it. The method of maximum likelihood has several peculiarities in its application to time-series analysis which are treated in this section. In particular we describe the conditional and unconditional likelihoods and sum of squares functions. We then derive the exact likelihood function (Gaussian case) of a moving average which in practice suffices for a general ARMA process.

Suppose we are given an ARMA model

$$(5.23) \quad e_t = X_t - p_1 X_{t-1} \ldots -p_n X_{t-n} + q_1 e_{t-1} + q_n e_{t-m}$$

Imagine for now that X_{-i} and e_{-i} the values of the observations, and the errors prior to the start of the series are known . Then using the known values for X_i the recursion can be solved to give the unobserved values of e_i. Thus, e_t is a function of P,Q,X_{-i} and e_{-i} which we denote by $e_t[P,Q|X_{-i}\ e_{-i}]$.

Since these values are iid normally distributed

$$(5.24) \quad p(e_1,\ldots,e_N) = a|\sigma|^{-N} \text{EXP}[-\frac{1}{2\sigma^2}(\sum_{t=1}^{N} e_t^2))]$$

where a is a real constant.

Thus we may write the likelihood conditional on the choice of X_{-i} and e_{-i} as above. The term in the exponential

$$(5.25) \qquad \bar{S}(P,Q) = \sum_{t=1}^{N} e_t^2[P,Q|X_{-i},e_{-i}]$$

where $P(B) = \sum_{i=1}^{n} p_i B^i$ and $Q(B) = \sum_{i=1}^{m} q_i B^i$

is called the conditional sum of squares function. The conditional likelihood is not yet a true likelihood as it contains unobserved values. In a long series however, this will make little difference because the inaccuracy in the choice of initial values will die out. However, since the likelihood is a useful practical tool we give the derivation of an exact likelihood function. This derivation for a moving average model will also furnish approximate likelihoods for mixed and autoregressive models on inversion and truncation of the operator series at convenient orders.

Thus consider the MA(m) model

$$(5.26) \qquad X_t = e_t + q_1 e_{t-1} + \ldots + q_n e_{t-m}$$

where the model is assumed to be invertible and e_{t+k} is independent of X_t for $k>0$. We derive the likelihood for a series of length n.

Because e_t and hence X_t are normally distributed we have

$$(5.27) \quad p([X]_{nx1}|Q,\sigma) = (2\pi\sigma^2)^{-n/2} |M|^{1/2} \text{EXP}[-\frac{[X]^T M [X]}{(2\sigma^2)}]$$

where $[X]_{nx1} = (X_1,\ldots,X_n)^T$

and $M^{-1}\sigma^2$ is the covariance matrix of [X]

Thus, to make this expression explicit one must evaluate $[X]^T_{nx1} M [X]_{nx1}$. To do this, first write the n+m equations

(5.28)
$$\begin{aligned}
e_{1-m} &= e_{1-m}\\
&\vdots\\
e_o &= e_o\\
e_1 &= X_1 + q_1 e_o + \ldots + q_m e_{2-m}\\
e_2 &= X_2 + q_2 e_1 + \ldots + q_m e_{2-m}\\
&\vdots\\
e_n &= X_n + q_n e_{n-1} + \ldots + q_m e_{n-m}
\end{aligned}$$

Now set

(5.29) $[e]_{m+nx1} = (e_{1-m},\ldots,e_n)^T$, $[\tilde{e}]_{mx1} = (e_{1-m},\ldots,e_0)^T$

Dropping the dimension subscripts, partition the resulting matrices as

(5.30)
$$[e] = \begin{bmatrix} 1\ 0\ 0 \ldots 0\ 0\ 0.. \\ 0\ 1\ 0 \ldots 0\ 0\ 0.. \\ 0\ 0\ 1 \ldots 0\ 0\ 0.. \\ \cdots\cdots\cdots\cdots\cdots \\ q_1\ q_2\ ..q_n \ldots 1\ 0\ 0.. \\ \quad .. \quad \ldots t_{21}\ 1\ 0.. \\ v_{21}\ \ v_{2n} \ldots t_{31}\ t_{32}\ 1.. \\ v_{n1}\ ..\ v_{nn} \ldots t_{n1}\ t_{n2}\ 1 \end{bmatrix} \begin{bmatrix} \\ [\tilde{e}] \\ \\ \cdots \\ \\ \\ \\ [X] \\ \\ \end{bmatrix}$$

$$= \begin{bmatrix} I & \vdots & O \\ \cdots & \vdots & \cdots \\ V & \vdots & T \end{bmatrix} \begin{bmatrix} [\tilde{e}] \\ \cdots \\ [X] \end{bmatrix}$$

where I is the identity, O is the zero matrix, T is lower triangular,and the coefficients of T and V are complicated functions of q_i obtained on expressing the e_i as functions of e_{-i} and [X]. It is seen from the first form that the transformation has unit jacobian, so the joint distribution of [X] and $[\tilde{e}]$ is

(5.31) $$p\{[X],[\tilde{e}]|Q,\sigma\} = (2\pi\sigma^2)^{-(n+m)/2} \; EXP\left[-\frac{1}{2\sigma^2} S(Q,[\hat{e}]) \right]$$

where, defining

$$K=\begin{bmatrix} I \\ - \\ V \end{bmatrix} \qquad L=\begin{bmatrix} O \\ - \\ T \end{bmatrix}$$

(5.32) $$S(Q,[\tilde{e}]) = \{ L\,[X] + K\,[\tilde{e}]\}^T \{ L\,[X] + K\,[\tilde{e}]\}$$

It is now sought to minimize $S(Q,[\tilde{e}])$ as a function of $[\tilde{e}]$. To do this write

(5.33) $$L\,[X] + K\,[\tilde{e}] = \{L\,[X] + K\,[\bar{e}]\} + \{K([\tilde{e}] - [\bar{e}])\}$$

so that if

(5.34) $$K^T K\,[\bar{e}] = K^T (L\,[X])$$

the quantities in braces are orthogonal vectors and so

$$(5.35)\quad S(Q,[\tilde{e}]) = S(Q) + \{[\tilde{e}] - [\bar{e}]\}^{T}\, K^{T} K \{ [\tilde{e}] - [\bar{e}]\}$$

and it follows $[\bar{e}]$ defined by the penultimate equation (5.34) is the minimizing value of $[\tilde{e}]$. Now by the above it follows that

$$(5.36)\qquad p\{[X],[\tilde{e}]|Q,\sigma\} =$$

$$(2\pi\sigma^{2})^{-(m+n)/2}\ EXP\left[-\frac{1}{2\sigma^{2}}\left(S(Q) + \{[\tilde{e}] - [\bar{e}]\}^{T} K^{T} K\{ [\tilde{e}] - [\bar{e}] \}\right)\right]$$

And since

$$(5.37)\qquad p\{[X],[\tilde{e}]|Q,\sigma\} = p\{[X]|Q,\sigma\}\ p\{[\tilde{e}]|[X],Q,\sigma\}$$

Conclude that

$$(5.38)\qquad p\{[\tilde{e}]|[X],Q,\sigma\} =$$

$$(2\sigma^{2})^{-m/2}\ |K^{T} K|^{1/2}\ EXP\left[-\frac{1}{2\sigma^{2}}([\tilde{e}]-[\bar{e}])^{T} K^{T} K\ ([\tilde{e}]-[\bar{e}])\right]$$

Therefore, on inspecting the exponential conclude

$$(5.39)\qquad [\bar{e}] = E\{ [\tilde{e}]|[X],Q \}$$

This means that $[\bar{e}]$, the minimizing estimate of $[\tilde{e}]$ may be computed from the difference equation by a technique known as backcasting. This is a trick which utilizes the fact that the distributions of $\mathbf{X}_{t}$ for the model written in forward form are the same as those for the model written in backward form, though as noted in section (1.4) the assumptions on the error sequence are different. The strategy is to compute the expected preliminary errors e_{-i} from the expected preliminary values of X_{-i}. Once the conditional expectations of the X_{-i} are known we

can combine this with knowledge that for a moving average of order m the conditional expectations $E\{e_{t-k} | X_t\} = 0$ for $k>m$. Using these facts in (5.26) we can obtain the conditional expectations for e_{t-k} $k=0,1-m$ given the data . Thus for example we compute (via 5.26) that $E[e_{-m} | \text{data}] = E[X_{-m} | \text{data}]$. If we denote conditional expectation given the data by square brackets this yields $[e_m] = [X_{-m}]$, $[e_{-m+1}] = [X_{-m+1}] - q_1 [e_{-m}]$, etc.

The trick is necessary because it will not do to compute the X_{-i} (or e_{-i}) by running the original recursion (5.26) backwards with estimated errors $\hat{e}_i$ for $i>0$. Because of invertibility, in this model inaccuracies die out with increasing time, but they must blow up as the time index decreases. A reversed version of this statement is true of the backward model

$$X_t = \epsilon_t + q_1 \epsilon_{t+1} + \ldots + q_n \epsilon_{t+m} \tag{5.40}$$

where ϵ_{t-k} is independent of X_t for $k>0$. That is, the inaccuracy in a guess for ϵ_t dies out as the time index decreases. Thus, using (5.40), accurate estimates of the ϵ_{-k} may be computed by guessing values near the end of the series and proceding backwards. These values of ϵ_t are used to construct the conditional expectations for X_{-i} $i>0$ in the

obvious way, and hence the expectations for e_{-k} may be found as described above. In this manner we arrive at

$$(5.41) \quad p\{[X]|Q,\sigma\} = (2\pi\sigma^2)^{-n/2} |K^T K|^{-1/2} \text{EXP}[-\frac{1}{2\sigma^2} S(Q)]$$

$$\text{where } S(Q) = \sum_{t=1-m}^{n} a_t^2$$

(a_t being conditional expectations of e_t given the data).

The necessary complicated function of the q_i may be found by iteration. Note however, that maximum likelihood estimates for time-series may not always be found by setting the derivatives of the log likelihood to zero, since the parameters corresponding to a given covariance structure are not unique. Thus it is necessary to constrain the maximization so that the resulting estimates satisfy the stationarity and invertibility conditions. We skip over the large literature on numerical computation of maximum likelihood and least squares estimates, because it is not relevant to the argument.

The purpose here is to make clear the conventional assumptions involved in the derivation and computation of maximum likelihood estimates for time-series. When the model is correct the likelihood function contains all useful information relevant to the estimation of parameters (cf. Box&Jenkins,1970). It appears that in addition likelihood functions may have a deeper significance than is evident from

this derivation. For even if false but close to correct (in some sense) we will see that the (false) likelihood contains valuable information. Thus, although the model is misspecified and the above theory is incorrect, a likelihood function may still be useful.

1.5.3 REVISING THE MODEL

It is possible and indeed likely that the identification stage of the time-series problem will have produced an incorrect choice of model. In other words, the application of the operator

$$Q_*(B)\, P_*^{-1}(B) \tag{5.42}$$

(where Q_* and P_* are estimated), to X_t may fail to transform the series to second order white noise (a sequence of uncorrelated random variables with zero mean). This outcome is likely first because the class of Box-Jenkins models is sufficiently rich that serious errors in model selection are probable. In addition, the requirement that a time-series be generated by a Box-Jenkins model is restrictive, and can be a misspecification in itself. When the parametrization is seriously wrong the maximum likelihood estimates will be invalid and we are in a situation which is at present, imperfectly understood.

There are two possibilities in this regard. The simplest occurs when it is possible to establish that the model is

incorrect through some diagnostic test. To cope with this situation, it is necessary to detect it, to elucidate the nature of the problem and modify the model appropriately. The other possibility is that several models may fit the data equally well, so that while identification has in a sense, failed, no obvious course of action is available. In practice it is usual to choose the simplest model which is acceptable.

To detect a failure in identification, two criteria have been used extensively in non-automatic procedures, namely, Quenouille's(1947) test and Box&Jenkins(1970) Portemanteau test.

If the $k+i^{th}$ partial autocorrelation is zero for i=0,m-k then (Anderson,1971) if $\hat{\phi}_l$ is the l^{th} estimated partial autocorrelation for the series, $\sqrt{N}\phi_{k+i}$ i=0,m-k have a limiting joint distributions which are independent N(0,1) under the null hypothesis. Thus, the statistic

$$Q = N \sum_{k+1}^{m} \hat{\phi}_i^2 \tag{5.43}$$

is distributed asymptotically as a chi-square random variable with m-k degrees of freedom. For an autoregressive order k gaussian model, the hypothesis that the model is correct is equivalent to the hypothesis that all partial autocorrelations after order k are zero. For white noise, this is true for k=0. Thus, this statistic may be applied to the residuals to detect failures of model fitting.

The Portemanteau test is an analogue of Quenouille's test with partial autocorrelations replaced by the usual autocorrelations Thus, if $\hat{e}_t$ are the residuals of the m^{th} order fitted model and $\rho_i^{\hat{e}}$ the i^{th} autocorrelation of the residuals, then the statistic

$$B_m = N \sum_{i=1}^{m} [\rho_i^{\hat{e}}]^2 \tag{5.44}$$

has an asymptotic chi-square distribution with m-p degrees of freedom.

These procedures for model selection terminate, in principle, when the residuals do not fail a test of randomness, and no simpler model has this property.

Several automatic and semi-automatic procedures for model determination have been proposed. The most well known of these is Akaike's (1973) criterion. This criterion is based on information theory and will be discussed in detail in Chapter III. Applied to the problem of choosing the correct autoregressive order, Akaike's criterion which would select an autoregressive order by minimizing

$$AIC(p) = \ln[\hat{\sigma}_p^2] + \frac{2p}{N} \tag{5.45}$$

where N is the sample size, $\hat{\sigma}_p^2$ is the estimate of the residual variance for the fitted model of order p.

When the hypothesis that the residuals are white noise is rejected, or if the model is unsatisfactory for some other reason, several kinds of situation may obtain. By linear model underfitting we refer to the fitting of a model with an insufficient number of parameters, that is models which set truely non-zero parameters to zero. Since it is usual to begin by fitting the simplest models this sort of problem should show up early in the fitting procedure, and be manifested in autocorrelated residuals and a large estimated error variance. Linear model overfitting refers to a situation in which a model of the correct type is fitted but the fitted model makes provision for non-zero values of parameters which are actually zero. The effect of this situation may be to reduce or increase the estimated prediction error variance and in a non-automatic procedure the only general means of testing for its occurence is to fit all simpler models to see whether they are acceptable. One case of overfitting, called parameter redundancy may however result in a certain pattern in the contours of the sum of squares (likelihood) function in the parameter space. For example consider the model

$$(1 - \alpha B)(1 - rB) X_t = (1 - sB) e_t \tag{5.46}$$

When $|r-s|$ is small the factors on each side of the difference equation will nearly cancel so that the model is nearly

$$(1 - \alpha B) X_t = e_t \tag{5.47}$$

As this is an operator equation, some discussion is required to establish that cancellation is meaningful. To see this, rewrite the model as

$$(5.48)\quad (1 - \alpha B)((1 - sB) + (s - r)B)X_t = (1 - sB)e_t$$

and invert explicitly to obtain

$$(5.49)\quad \{1 + (s-r)B\{1 + sB + s^2B^2 + \ldots\}\}(1 - \alpha B)X_t = e_t$$

This means that the behavior of the series with nearly equal values of s-r will ressemble one another. In particular, changes in (s,r) in the neighbourhood of the line s=r will have little effect on the goodness of fit of the model. Thus if the true parameters r and s are nearly one it will be difficult to discriminate a stationary from a non-stationary model when the more complex one is fitted. However examination of the likelihood will reveal a ridge in the sum of squares function along the line s=r, indicating parameter redundancy. At the expense of a small increase in lack of fit, the offending factors may be dropped from the model. This then, is a practical situation in which the nature of misspecification is evident from the likelihood function. We will see later that other, more theoretical, approaches to the detection of misspecification are also based upon likelihood functions.

1.6 USE OF RESIDUALS TO MODIFY THE MODEL

Suppose that the residuals $\hat{\epsilon}_t$ from the fitted model

$$P(B)X_t = Q(B)\epsilon_t \quad (5.50)$$

appear to be non-random in the sense of section 1.5.3. Suppose further that one could correctly fit the model

$$R(B)\epsilon_t = S(B)e_t \quad (5.51)$$

where e_t are iid random variables with zero mean and the model defined by R and S is stationary and invertible.

Then, the model could be corrected to

$$P(B)X_t = Q(B)\{R^{-1}(B)S(B)\}e_t \quad (5.52)$$

or, avoiding infinite series

$$P(B)R(B)X_t = Q(B)S(B)e_t$$

1.7 SUMMARY

The time-domain theory for linear time-series was reviewed, including the relationship between covariance structure and Box-Jenkins coefficients, and the interpretation of the invertibility and stationarity conditions. Current methods of attacking the linear time-series problem were reviewed. It was seen that these procedures may require judgement in detecting misspecification. There is some indication, however that the likelihood function contains much of the relevant information

for this judgement. Thus, it is natural to ask whether automatic procedures based on the likelihood function are a practical way to address the problem of misspecification. If such automatic procedures were marked improvements on the skilled procedures, they would be helpful in the (harder) non-linear case. Succeeding sections will evaluate this possibility.

II. HOW WELL DO THESE METHODS WORK?

The paper of Bora-Senta& Kounias(1979) systematically compares order determination schemes using Quenouille's test, the Portemanteau test and Akaike's criterion for various estimation methods in a Monte Carlo simulation. We present some of the results of the study in Tables I-III. The method of estimation used was the method of unconditional least squares. Significance levels for the portemanteau test were determined from the theoretical Chi-squared variate distributed with m-p degrees of freedom where m=10 and p is the order of the fitted model. Similarly, significance levels for Quenouille's test were determined from the Chi-squared variate distributed with m-p degrees of freedom with m=6 and p the order of the fitted model. Thus the comparison between Portemanteau and Quenouilles test may not be fair, though neither test can be valid owing to the small sample size (n=20). The model adopted was the model of lowest order which passed the respective tests at the .05 level of significance.

All simulations were involved one of the 6 autoregressive models quoted in Tables I-III. For a large sample size (n=200) fifty simulation runs were performed and all criteria performed well. Quenouille's test was the best performer, failing in only 1/6 cases to select the correct AR order on all fifty runs. The other two criteria, each failed in 2/6 cases to have a perfect record, misclassifying at least 10 runs when they failed.

For the small sample size, performance is less impressive.

Examination of the modal choice of order shows only Quenouilles test has the wrong mode (AR(5)) for the first model. Both portemanteau and Quenouille are wrong for the second AR(1) model, though only Portemanteau is wrong if the choice AR(5) is disallowed. All criteria fail to have correct modal choice of order for the first AR(2) model, each criterion instead selecting the AR(1) model most frequently. For the second AR(2) model correct modes resulted with each criterion.

Performance was further summarized by the use of the index W

$$W(p_0) = \sum_{\hat{p}=0}^{5} (p_0 - \hat{p})^2 f_{\hat{p}} / N \qquad (1.1)$$

where p_0 is the true order, and $f_{\hat{p}}$ is the frequency of selection of order $\hat{p}$.

By this index, Portemanteau wins for the first and second AR(1) models with Akaike a close second. In the other 4/6 cases Akaike wins, decisively in 3/6 cases. Mean values of W are respectively 1.398, 5.607, 2.147 for Akaike, Quenouille and Portemanteau respectively. The relatively poor performance of Portemanteau and Quenouille might be expected here since as noted they are large sample tests, however, it should also be pointed out that the order selection problem considered by Bora-Senta et. al. was a simplified one in which the alternatives were just six autoregressive models (order=0,..,5) which were assumed a priori to be stationary.

When models are allowed to be non-stationary by introducing

powers of difference operators into the defining equations and moving average terms, then the model-discriminating powers of these procedures begins to be taxed even at large sample sizes. Ozaki(1977) compared the models of data series A-F (data given in Box&Jenkins book) obtained by Box&Jenkins with those adopted by Akaike's criterion. We discuss his results for series A, which are the most completely presented.

Series A from Box&Jenkins(1970) consists of 197 observations of concentration readings for a chemical process taken every two hours. From inspection of the sample autocorrelation function and partial autocorrelation function Box&Jenkins suggest that this time-series might be described by an IMA(0,1,1) or ARMA(1,0,1) model. Ozaki concluded however that (7,0,0) and (6,1,0) and (1,1,1) were also candidate models, by the same procedure.

The results of fitting these models to the data are given in Table IV. All these models passed the Portemanteau test.

Akaikes criterion was useful in automatically weeding out a large number of other models (which likely would have failed the eyeball test) but somewhat less helpful in discriminating amongst models which were good candidates. Values of Akaike's criterion for the various models considered are given in Table V. While Box&Jenkins concluded that each of (1,0,1) and (0,1,1) fitted equally well Akaikes criterion gives (1,1,1) as the model of choice amongst

(1.2) (i) $\{ (p,d,q),\ d=0,1 \quad 0 \leq p,q \leq 2 \}$

and (3,0,3) as the model of choice amongst

(ii) $\{ (p,d,q),\ d=0,1 \quad 0 \leq p,q \leq 9 \}$

However, the (3,0,3) model failed the Portemanteau test.

It can be seen that the models adopted by Box&Jenkins exhibit AIC values near the minimum. On the other hand, while Ozaki claims that the automatic procedure's results are similar to those of Box &Jenkins, examination of the numbers seems to show that Akaike's procedure recomends 9 of a possible 18 models (i) as well as recommending the higher order models of Table IV. In particular 9 models had AIC values between -449 and -451 whereas, for comparison, the models selected by Box&Jenkins had AIC(1,0,1)=-450.2 and AIC(0,1,1)=-448.16. See Table V. Now it is very well to say that one should apply Ockham's razor but in the first place, 3 of these models have the same number of parameters, and in the second, it is not sensible to do this without regard for the explanatory power of the model . Akaike's criterion, which is supposed to formalize the tradeoff between explanatory power and parsimony is what gives us these results in the first place. Thus the data seem ambiguous. Examination of the other results in Ozaki 's paper suggest that this situation is not untypical. This problem is less disastrous than it appears, since in several cases the different models differ only by the addition of nearly cancelling factors on each side of the equation, which implies that very similar forecast would be obtained from these apparently different

models.

The conclusion appears to be that while practice shows that useful prediction can be obtained, optimal prediction may be difficult, for presumably not all of the models the data admits will be correct. Automatic procedures give a clear reduction in labour, though not necessarily an improvement in performance. Ozaki remarks that a classical Box-Jenkins identification by visual inspection of the autocorrelations etc, can produce reasonable results, but only by 'skilful analysis which requires expertise perhaps not commonly possessed by ordinary practicing statisticians.' Both automatic and skilled procedures invariably result in the acceptance of some useful model. However, the results of either procedure are usually far from unique when the class of admissible models is large.

Table I - AR order Selection

Model	$X(t)= .3\ X(t-1) + e(t)$		
order selected	Akaike	Quenouille	Portemanteau
0	360	140	480
1	60	10	10
2	30	0	10
3	40	30	0
4	10	0	0
5	0	320	0
W	1.28	10.76	.98

Model	$X(t)= .8\ X(t-1) + e(t)$		
0	70	40	310
1	300	120	180
2	70	10	10
3	50	10	0
4	0	30	0
5	10	290	0
W	1.0	10.6	.64

Method of maximum likelihood was used for estimation of coefficients. Sample series were of length 20.

Table II - AR order selection

Model X(t)= .3 X(t-1) + .5 X(t-2) + e(t)

order selected	Akaike	Quenouille	Portmanteau
0	240	145	410
1	105	55	90
2	110	90	0
3	35	45	0
4	5	45	0
5	5	210	0
W	2.33	5.5	3.28

Model X(t)= .8 X(t-1) -.9 X(t-2) + e(t)

0	0	0	90
1	0	0	0
2	400	300	410
3	70	40	0
4	20	30	0
5	10	130	0
W	.48	2.6	.72

Table III - AR Order Selection

Model $X(t) = 1.6\,X(t-1) - .79\,X(t-2) + .12\,X(t-3) + e(t)$

order selected	Akaike	Quenouille	Portemanteau
0	0	0	40
1	110	30	380
2	320	230	80
3	60	20	0
4	10	10	0
5	0	210	0
W	1.54	2.4	3.76

Model $X(t) = 1.8\,X(t-1) - 1.14\,X(t-2) + .272\,X(t-3) + e(t)$

0	0	0	40
1	130	20	310
2	320	255	150
3	25	60	0
4	20	35	0
5	5	130	0
W	1.76	1.78	3.50

Table IV - Models for Series A

Model	N-d	Fitted Model	Residual Variance	Portemanteau Statistic	d
(7,0,0)	197	$1-.355B -.186B^2 -.023B^3-.029B^4 -.021B^5-.085B^6 -.203B^7$	.0925	13.61	18
(6,1,0)	196	$1+.624B +.423B^2 +.386B^3 +.339B^4 +.346B^5 +.238B^6$	.0945	15.05	19
(1,1,1)	196	$(1-.218B)(1-B)X(t) = (1-.825B)\ e(t)$	.0985	23.37	
(1,0,1)	197	$(1-.915B)X(t) = (1-.583B)e(t)$	.0977	26.45	
(0,1,1)	196	$(1-B)X(t) = (1-.705B)e(t)$	.1007	29.87	

All the above models passed the portemanteau test with p=.05, k=25 (d.f.=k-p-q for model (p,d,q)).

Table V - Values of AIC for Various Models

Model	AIC	Model	AIC
(1,0,1)	-450.26	(1,1,1)	-450.55
(2,0,1)	-450.26	(2,1,1)	-449.72
(1,0,2)	-449.73	(0,1,2)	-449.53
(2,0,2)	-448.67	(1,1,2)	-449.23
(2,0,0)	-445.73	(0,1,1)	-448.16
(1,0,0)	-434.58	(2,1,2)	-447.47
(0,0,2)	-423.76	(2,1,0)	-430.56
(0,0,1)	-403.08	(1,1,0)	-425.57
(0,0,0)	-358.76	(0,1,0)	-390.40
*(3,0,3)	-456.16		
*(7,0,0)	-450.89		
*(6,1,0)	-450.68		

III. METHODS OF SELECTING AR ORDER

Statisticians are often faced with the problem of choosing the appropriate dimensionality of a model that will fit a given set of observations. The problem is particularly acute in the analysis of time-series data. This is because, in practice, the time-series problem is often posed to the statistician in the form we have posed it, without much knowledge of the underlying structure. In this situation, the problem must be attacked frontally despite its formidable nature. Thus, whether the aim is structural characterization or optimal prediction, the parametrization (misspecification) problem in time-series analysis is basic. It is the purpose of this section to describe some recent work on this problem.

It is part of statistical folklore (cf. Schwartz,1978) that the method of maximum likelihood performs badly in the selection of over-parametrized models. In particular there is a bias toward the selection of models with too high a dimensionality. With the method of least squares for example, the minimization of prediction error on the data set leads to non-zero estimates of parameters which are zero and biases the estimate of mean square error.

Quenouilles test and the Portemanteau test were derived with no particular alternative in mind and in practice are applied repeatedly (which is not a recommended approach to hypothesis testing) however they represent a considerable improvement on simple maximum likelihood. In particular these

tests may be understood as likelihood ratio tests against particular alternatives so that they are maximum power tests under the usual conditions(Hosking,1978). Thus,the use of the Portemanteau test statistic

$$N \sum_{k=1}^{m} \hat{\rho}_k^2 \sim \chi^2_{m-p} \tag{1.1}$$

is equivalent to a likelihood ratio test of

$$H_0 : P(B)Y_t = a_t \tag{1.2}$$

versus

$$H_1 : T(B)P(B)Y_t = \mathring{a}_t$$

where a_t and $\mathring{a}_t$ are white noise processes possibly with different variances. Thus the alternative is an autoregression fitted to the residuals or from another point of view, an autoregression of order m+p with p roots restricted to the same values as the null hypothesis.

Quenouille's test statistic (Whittle,1952)

$$N \sum_{k+1}^{m} \hat{\phi}_i^2 \sim \chi^2_{m-k} \tag{1.3}$$

is equivalent asymptotically to a likelihood ratio test of

$$H_0 : P(B)Y_t = a_t \tag{1.4}$$

versus

$$H_1 : S(B)Y_t = a_t$$

where deg P = p and deg S = p+m for small variations about the null hypothesis (the size of the variation affects the approximations used in deriving the results).

As we noted, Quenouille's test and the Portemanteau test

are usually applied repeatedly. Therefore this approach is not completely satisfactory. Another set of approaches may be thought of as based on the following empirical observation, well known to statisticians . Namely, that if spurious explanatory variables are added to a regression model, estimated mean square error first exhibits a drop as the first few are added, then after a certain number have been added, mean square error begins to rise. This suggests that some function of estimated mean square error might be a sensible order selection criterion. Note that in the Gaussian case the mean square error is an estimate of log likelihood. There have been a number of attempts to formalize such procedures of which we will discuss the most systematic.

To illustrate the phenomenon we mention the simulation study of Jones(1976). In this study, autoregressive models of various orders were fitted to white noise for samples of size 20 and 40 and the results averaged over 100 realizations. Possibly because white noise was being fitted, no initial drop in mean square error was evident as the order of the autoregression was increased. However, estimated mean square error rose drastically as the order, p of the fitted autoregression approached the length of the sample. Thus, estimated error doubled as p approached 15/20 and 25/40 respectively.

Akaike's(1969) final prediction error criterion was based directly on this phenomenon. It is

$$\text{FPE}(p) = \hat{\sigma}^2_p \left(1 + \frac{(p+1)}{N}\right) \tag{1.5}$$

where

$$\hat{\sigma}_p^2 = \sum_t^N (\hat{e}_t^{[p]})^2 / (N - 1 - p)$$

and $\hat{e}_t^{[p]}$ denote the residuals from the fitted autoregression of order p. The factor multiplying the mean square error is calculated to compensate for the effect of errors in estimating the parameters so that FPE is an estimate of the average prediction error when a fitted model of order p is used. Jones(1975)and Gersh&Sharp(1973) presented favourable simulation and practical results for the use of this criterion in the selection of autoregressive order. Minimization of final prediction error turns out to be a special case of minimization of Akaike's(1973) information criterion which, because of its widespread use and interesting theoretical rationale, we discuss in some detail.

Akaike's information criterion is

(1.6) -2log likelihood + 2 (no. of parameters adjusted)

Heuristically, the idea is to measure the amount of structure in the estimated model (ie. in its probability density) and to choose the fitted model of a class which exhibits the least structure. Such a procedure, made precise, is conservative in the sense that it attributes the least possible structure to the data, consistent with membership in a certain class of models. It will turn out that the definition of structure employed will assign less structure to models with small error variance, and

to models with small numbers of parameters. While the notion that more parameters mean more structure should cause no difficulty, that a small variance should correspond to less structure may seem problematic. It is useful to think of the case of a discrete random variable in which a large variance might mean that more different, widely scattered values occur in the distribution. Alternatively Akaike's criterion can be thought of as one possible choice of loss function in the trade-off between reduced (sample) error variance and number of parameters, that is, a formalization of the principle of parsimony. In the case of a one parameter family of distributions, Akaike's method obviously reduces to the method of maximum likelihood.

We now present a series of heuristic arguments justifying Akaike's criterion, beginning from the desire to develop a model selection procedure which is conservative in the above sense. To begin, we argue that a parameter in a model represents structure only because it indexes significant deviations from some reference model. So it is reasonable to think of small deviations from reference as implying less structure, and larger deviations as implying more structure. It is natural to take the true model as reference, so that the attribution of least structure leads to a correct choice when the model is known. These considerations mean that we need to look for an index of the deviations of a putative model from the true one.

The well known fact that for large N the maximum likelihood

estimator approaches that theoretical minimum variance estimator suggests that the likelihood function is a sensible index of deviation of the model parameters from the true values. This motivates the assumption that the information (structure) function I is proportional to

$$(1.7) \quad -I\{f,\Phi\} = \int \Phi\{ f(x|\theta)/ f(x|\theta_o) \} f(x|\theta_o)\, dx$$

$$= E\{ \Phi\{ f(\theta)/f(\theta_o) \} \}$$

where f is the likelihood function of the observations. The choice of Φ is equivalent to the choice of a loss function for decision making, for if the loss function were chosen as $\lambda\{\theta,\theta_o\}$ and it were sought to minimize $E[\lambda[\theta,\theta_o]\}$ then this expectation could be defined as the previous integral for some Φ provided that

$$(1.8) \quad \frac{\partial f}{\partial \theta} \neq 0$$

using the inverse function theorem. The choice of the appropriate Φ follows from the hypothesis that, if X and Y are independent random variables then

$$(1.9) \quad I(X + Y) = I(X) + I(Y)$$

where I(.) is the information function

Since the joint density (X,Y) is the product of the densities of X and Y, it immediately follows that the appropriate choice of Φ is

$$(1.10) \quad \Phi\{r\} = c \log(r)$$

and furthermore, if f and g are probability densities

$$E_g \{ \log f \} = \max \tag{1.11}$$

iff f=g a.e. since this is true when f and g are simple functions, using a lagrange multiplier argument.

In practice, the true density will be unknown, so that we will work with the natural estimate of $\Phi(\theta_o)$ namely,

$$\frac{1}{N} \sum_i \log f(x_i \mid \hat{\theta}) \tag{1.12}$$

It is now possible to formulate the problem of order estimation precisely. Suppose then that we want to choose a density $f\{x|\theta\}$ which best approximates the density $f\{ x|\theta_o \}$ for $\theta_o \epsilon \Theta_k$ and $\theta \epsilon \Theta_k$ where Θ_k is a k dimensional subspace of Θ. Thus, the problem of order choice is replaced with a problem in density estimation. We define $\hat{\theta}$ and $\hat{\theta}^k$ to be the maximum likelihood estimates of θ on restriction to the respective linear spaces. Note the notational distinction between $\hat{\theta}_k$ which is the maximum likelihood estimate of θ_o projected onto Θ_k, and $\hat{\theta}^k$ which is the maximum likelihood estimate in the restricted parameter space Θ_k. Now the ultimate object is to obtain the estimate $\underline{\theta}$ which minimizes $I(f(\theta))$. However this is not completely straightforward. If k is fixed and $\theta_o \epsilon \Theta_k$ then the

result of maximizing the function

$$F(\theta) = \frac{-2}{N} \sum_i \log f\{x_i \mid \theta\} \tag{1.13}$$

will be just the ordinary maximum likelihood estimator. However when $\theta_o \notin \Theta_k$ the statistic $F(\hat{\theta}^k)$ cannot estimate $I(f(\theta))$ or $_o$ a fortiori $I(f(\hat{\theta}))$ for $\hat{\theta} \neq \theta_o$ and some adjustment becomes necessary. This is done as follows: define the matrix

$$J_{ij} = E\left[\frac{\partial \log f\{X \mid \theta\}}{\partial \theta[i]} \cdot \frac{\partial \log f\{X \mid \theta\}}{\partial \theta[j]} \right]\Big|_{\theta_o} \tag{1.14}$$

$$= -E\left[\frac{\partial^2}{\partial \theta[l] \partial \theta[m]} f\{X|\theta\} \right]\Big|_{\theta_o}$$

for i,j=1,...L where L is the dimension of Θ. This is the Fisher information matrix.

It is ordinarily positive definite, being the variance covariance matrix of the random variable(s)

$$\frac{\partial f\{\theta\}}{\partial \theta[i]} \frac{1}{f\{\theta\}} \tag{1.15}$$

Let $\underline{\theta}$ be the maximizing argument of $I(\theta_o, \theta)$ restricted to Θ_k.

Finally define the J inner product as $\{u,v\} = u^T J v$ then

(i) it can be shown that approximately

$$I(\theta_o, v) = \frac{-1}{2} \| \theta_o - v \|_J^2 \tag{1.16}$$

and

(ii) $\underline{\theta}$ is approximately the projection of θ_o on Θ_k in the J inner

product. Thus (approximately) by the pythagorean theorem

$$(1.17) \qquad || \underset{o}{\theta} - \hat{\theta}^{k} ||^{2}_{J} = || \underline{\theta} - \underset{o}{\theta} ||^{2}_{J} + || \hat{\theta}^{k} - \underline{\theta} ||^{2}_{J}$$

and thus

$$(1.18) \qquad -I(\underset{o}{\theta}, \hat{\theta}^{k}) = -I(\underset{o}{\theta}, \underline{\theta}) + N|| \hat{\theta}^{k} - \underline{\theta} ||^{2}_{J}$$

We want to estimate $I(\underset{o}{\theta} ,\underline{\theta})$, but can only hope to observe a function equal to a constant (relative to k) plus

$$(1.19) \qquad N\, I(\underset{o}{\theta} , \hat{\theta}^{k})$$

$$= -[\Sigma \log f\{x_i | \hat{\theta}^{k}\} - \log f\{x_i | \underset{o}{\theta}\}]$$

on restriction to each subspace. However, it can be shown (iii) that asymptotically

$$(1.20) \qquad N|| \hat{\theta}^{k} - \underline{\theta} ||^{2}_{J} \sim \chi^{2}_{k}$$

So that the statistic (1.19) can be made unbiased (up to a constant in k) for $I(\theta,\underset{o}{\theta})$ by adding $E\,\chi^{2}_{k}$. Since k is just the number of independently adjusted parameters this reasoning leads to Akaike's statistic.

Following Akaike(1973), we now sketch the proof of the necessary facts. Approximately here means ignoring terms of order $|| \underset{o}{\theta} - v ||^{3}$. Fact (i) follows from a Taylor expansion

of $I(\theta,v)$ assuming that
$\quad\ \ o$

$$(1.21) \qquad \frac{\partial I(\theta_o,\theta)}{\partial\theta}\Big|_{\theta=\theta_o} = 0$$

Fact (ii) follows from the first fact and the variational characterization of projections. To get the third fact procede as follows. Assume that Θ contains $\underset{o}{\theta}$ and recall $\hat{\theta}$ denotes the unrestricted maximum likelihood estimate of $\underset{o}{\theta}$. Assume

$$(1.22) \qquad \sum_i \frac{\partial \log f(x[i]|\hat{\theta})}{\partial\theta[m]} = 0 \qquad \text{for } m=1,\ldots,L$$

$$(1.23) \qquad \sum_i \frac{\partial \log f(x[i]|\hat{\theta}^k)}{\partial\theta[m]} = 0 \qquad \text{for } m=1,\ldots,k$$

Then using the mean value theorem

$$(1.24) \qquad \sum_{i=1}^{N} \log f(x_i|\underline{\theta})$$

$$= \sum_{i=1}^{N} \log f(x_i|\hat{\theta}) + \frac{1}{2}\sum_{m=1}^{L}\sum_{l=1}^{L} N(\underline{\theta}[m]-\hat{\theta}[m])(\underline{\theta}[l]-\hat{\theta}[l]) \times$$

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial^2}{\partial\theta[m]\partial\theta[l]} \log f(x_i|\hat{\theta}-\rho(\underline{\theta}-\hat{\theta}))$$

$$= \sum_{i=1}^{N} \log f(x_i|\hat{\theta}^k) + \frac{1}{2}\sum_{m=1}^{k}\sum_{l=1}^{k} N(\underline{\theta}[m]-\hat{\theta}^k[m])(\underline{\theta}[l]-\hat{\theta}^k[l]) \times$$

$$\frac{1}{N}\sum_{k=1}^{N} \frac{\partial^2}{\partial\theta[m]\partial\theta[l]} \log f(x_i|\hat{\theta}^k-\rho^k(\underline{\theta}-\hat{\theta}^k))$$

where $0 \le \rho^k, \rho \le 1$.

We assume that the matrices in the second terms converge in probability as $N \to \infty$ to the corresponding Fisher Information matrices.

On differentiating the above expressions with respect to $\underline{\theta}$ and dividing by N it is seen that for $l=1,k$

$$(1.25) \qquad \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial\theta[l]} \log f(x_i \mid \underline{\theta})$$

$$= \sum_{m=1}^{k} (\underline{\theta}[m] - \hat{\theta}^k[m]) \frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \log f(x_i \mid \hat{\theta}^k + \rho(\underline{\theta} - \hat{\theta}^k))}{\partial\theta[m]\partial\theta[l]}$$

$$= \sum_{m=1}^{L} (\underline{\theta}[m] - \hat{\theta}^k[m]) \frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \log f(x_i \mid \hat{\theta} + \rho(\underline{\theta} - \hat{\theta}))}{\partial\theta[m]\partial\theta[l]}$$

We need the first member in this sequence of equations to tend to zero as $N \to \infty$. This is reasonable given some mild conditions on f since $\underline{\theta}$ maximizes the structure function I by hypothesis. It would hold for example if $\underline{\theta}$ is a normal mean or variance. If so then the last equations imply that asymptotically (for N large)

$$(1.26) \qquad J_{kxk} (\underline{\theta} - \hat{\theta}^k) = J_{Lxk} (\underline{\theta} - \hat{\theta}^k)$$

or in other words

$$(1.27) \qquad \Pi\, J_{LxL}\, \Pi(\underline{\theta} - \hat{\theta}^k) = J_{LxL}\, \Pi(\underline{\theta} - \hat{\theta})$$

where Π denotes the J-orthogonal projection onto Θ_k, J is the Fisher Information matrix.

Since the equation (1.27) characterizes the projection we conclude that asymptotically

$$\Pi(\underline{\theta}-\hat{\theta}) = \underline{\theta}-\hat{\theta}^{k} \tag{1.28}$$

which may be written

$$(\underline{\theta}-\hat{\theta}^{k}) = (\underline{\theta}-\hat{\theta}_{k}) \tag{1.29}$$

We can now use this equality to derive the distribution of $N||\hat{\theta}^{k}-\underline{\theta}||^{2}_{J}$ from the assumption that $N^{1/2}(\hat{\theta}-\theta) \xrightarrow{D}$ $N(\underline{0},J^{-1})$. First note that for any Y in Θ, if $Y \sim MVN(0_{1},K^{-1})$ (multivariate normal) this implies there exists a matrix M such that that MY is standard normal of dimension 1 and $M^{T}M = K$. Since

$$||Y||^{2}_{K} = Y^{T}KY = Y^{T}M^{T}MY \tag{1.30}$$

it follows the norm (1.30) is χ^{2}_{1}. Thus, we need only show that in a space of dimension L, if Π is the J-orthogonal projection (for J positive definite) onto a subspace Θ_{k} of dimension $k < L$, and $X \sim MVN(0_{L},J^{-1})$ then

$$\Pi X \text{ is } MVN(0_{k},J^{-1}|_{\Theta}). \tag{1.31}$$

This is certainly true if J is diagonal, since the density of an independent MVN distribution will factor. Let P diagonalize J. Then in P coordinates the same result holds. This is equivalent to the projection result (5.31). Finally because J is positive definite the L-dimensional distribution, and thus the k-dimensional distribution is non-degenerate. Thus we conclude that $||\underline{\theta} - \hat{\theta}^{k}||_{J}$ is distributed as χ^{2}_{k} as required.

Another technique which was formulated to solve the parametrization problem is known as cross validation. It is based on the notion of optimizing final prediction error estimated in a slightly different manner. The procedure is as follows for each candidate parametrization. For each data point $\{ x_i, y_i \}$ compute the maximum likelihood estimate for the parameters of interest with this data point omitted. Thus compute $\hat{\theta}_{-i}$ for the omitted ith point. Measure the prediction error by the squared difference from the observed value, or more generally $\log f((x_i, y_i)|\hat{\theta}_{-i})$. Repeat for each sample point and add prediction errors. Chose the parametrization giving the least summed error. Stone(1977) shows that this technique is asymptotically equivalent to Akaike's criterion under assumptions similar to those above. Shibata (1976) showed that Akaike's technique is not consistent, and that it over-estimates autoregressive order with non-zero probability as N --> ∞. This implies that cross-validation also

fails to be a consistent technique. Hannan(1979) has modified Akaike's criterion so that the result is consistent.

IV. STRUCTURAL PROPERTIES: LINEAR EXAMPLES

We have seen that data may be ambiguous with respect to the correct choice of parametrization. This suggests that it will be advantageous to try to employ any available structural information even though it may not be in convenient form. It will be suggested that recent proposals for non-linear models ought to be examined with this emphasis rather than as general data prescriptions. That is, in contrast to the widely followed emphasis of Box&Jenkins on modelling data for which little or no a priori structure is postulated, we suggest that time-series data analytic models must be chosen in accord with known general facts about the natural process. Succeeding sections will deal with how (and why) non-linear models may be utilized for this purpose.

We first show that within the class of linear models it is possible to augment the useful information available by an appeal to relatively vague structural knowledge. The fact that these examples due to Granger& Morris(1976) are not better known illustrates the neglect of structural data by time-series analysts. Suppose it is reasonable to assume that an underlying time-series X_t is Markov. In the linear case this means that the partial autocorrelation function cuts off and $\phi_i = 0$ for $i>1$, that is no additional useful information for prediction exists once X_{t-1} is known. This might be the case for instance

in a physical model for which all theoretical models are differential equations, and so are all Markov. Suppose however that an independent source of white noise is added to the observations.

The time-series described above will be identified by Box-Jenkins procedures as an ARMA(1,1) process. To see this define

$$(1.1) \qquad (1 + aB)\, X_t = w_t$$

$$Y_t = v_t$$

where

a is a real constant such that $|a| \leq 1$, B is the backward shift operator, and w_t and v_t are uncorrelated white noise processes.

Suppose that we are observing

$$(1.2) \qquad Z_t = Y_t + X_t$$

then:

$$(1.3) \qquad (1+ aB)\, Z_t = (1 + aB)\, v_t + w_t \qquad (= m_t)$$

Computing the autocorrelations for the process on the right hand side $(=m_t)$ of the latter equation we get:

$$(1.4) \qquad u_0 = E(m_t^2) = (1 + a^2)\, \sigma_v^2 + \sigma_w^2$$

$$u_1 = E(m_t m_{t-1}) = a\sigma_v^2$$

and

$$E(m_t m_{t-i}) = 0 \qquad \text{for } i \geq 2$$

We will construct a putative representation in the form

(1.5) $$(1 + c B) Z_t = (1 + d B) e_t \qquad (=n_t)$$

by choosing the parameters c,d and σ_e^2 to produce the values of u_i computed above . To this end we let c=a, and define the other parameters via the equations

(1.6) $$u_1 = d \sigma_e^2$$

and

$$u_o = (1 + d^2) \sigma_e^2 \quad .$$

to complete the construction. However, observe that these equations can always be solved, and that the resulting processes (n_t) have the same covariance structure as (m_t). Thus the construction works and generates a process with the same covariances as (1.2).

In contrast, it may not be possible to represent ARMA(1,1) as AR(1)+ white noise . The realizability (necessary) conditions may be stated as follows.

Let the u_i be given by the ARMA(1,1) process (1.5), as computed above . Define

(1.7) $$p_1 = u_1 / u_o$$

Then

$$(1.8) \qquad p_1 \equiv a^2\sigma_v^2 / ((1 + a^2)\sigma_v^2 + \sigma_w^2)$$

$$\leq a^2/(1 + a^2)$$

$$= c^2/(1 + c^2) \ .$$

As long as these inequalities are satisfied, some choice the parameters σ_v^2 and σ_w^2 will allow the construction of a candidate AR(1)+noise model. Therefore (1.8) is a realizability condition on the ARMA process for an AR(1)+noise representation.

Thus, under the original assumptions, this condition should be satisfied by parameter estimates for AR(1)+noise and its application will simplify computation of maximum likelihood estimates. It is possible moreover, to reduce the number of parameters in some cases. For example, using methods like these it can be shown that AR(2)+noise identifies as ARMA(2,2). Writing out the respective models it may be seen that ARMA(2,2) involves five parameters and AR(2)+noise only four. For if

$$(1.9) \qquad Z_t = X_t + Y_t$$

where $(1 + a_1 B + a_2 B^2)\, X_t = \epsilon_t$ and $Y_t = \eta_t$

then

$$(1.10) \qquad (1 + a_1 B + a_2 B^2)\, Z_t = (1 + a_1 B + a_2 B^2)\, \eta_t + \epsilon_t$$

where ϵ_t and η_t are independent white noise processes, while the

ARMA model is

$$(1.11) \qquad (1 + c_1 B + c_2 B^2) Z_t = (1 + d_1 B + d_2 B^2) \zeta_t$$

where ζ_t is a white noise series.

This fact is reflected in the realizability conditions for representing ARMA(2,2) as AR(2)+noise . As before let the symbols u_i denote the successive autocorrelations of the process on the right hand side of the above equations. Define

$$(1.12) \qquad p_1 = u_1 / u_o$$

and

$$p_2 = u_2 / u_o$$

It may be shown that a given ARMA(2,2) process is expressible as AR(2)+noise provided

$$(1.13) \qquad c_2 / (1 + c_1^2 + c_2^2) > p_2 \geq 0$$

and

$$c_2 / c_1 (1 + c_2) = p_2 / p_1$$

Because these conditions involve an equality, the dimension of the parameter space has been reduced. This is a gain in simplicity which, if justified, will reduce the final error of prediction.

V. REPRESENTATION THEORY FOR WEAKLY STATIONARY PROCESSES:IMPLICATIONS FOR NON-LINEAR TIME-SERIES MODELLING

5.1 WOLD'S THEOREM AND SINGULAR PARTS

We contend that to make full use of non-linear time-series models the use of available structural information is essential [1] . The insistence on structural restrictions is closely related to prediction since enlarging the class of admissible models means that the errors made in choice of parametrization become more important. These errors affect the final prediction error. It seems clear from our discussion of AR order choice that this phenomenon leads to difficulties in the case of linear identification, and that in the non-linear case the problem will be worse.

This observation is of little use without some clarification however, since the notion of structure is multifaceted. To the scientific investigator, structural models are those which express interpretable relationships between a

---

[1] Some non-linear time-series modelling might be possible with minimal resort to structural information if a useful non-parametric approach could be formulated. There is already a literature on non-parametric regression (eg. Stone,1977; Friedman& Stuetzle,1981). To make this work for time-series one might start with an extremely general autoregressive model. For example, a time-series might be modelled as the solution of an autonomous continuous time stochastic differential equation (SDE) (cf. Jones,1981; Ludwig,1975) and the parameters estimated locally using techniques of non-parametric regression. The special case when the underlying deterministic flow is a gradient field was investigated and partially solved, in work related to this thesis. Unfortunately this investigation had to be abandonned due to time constraints.

priori state variables in the system of interest. To the statistician structure means the specification of a data prescription. In order to reconcile these views we need to formulate the idea of the structure of a process in a more precise fashion.

Recall that our definition of a time-series explicitly mentions a deterministic relation forced by errors, so that conceivably, not all weakly stationary processes may correspond to time-series. This section is devoted to clarifying the probabilistic structure of weakly stationary processes and their representations as time-series . From this discussion conclusions are drawn about the connection between interpretable relationships and data prescriptions, as well as the appropriateness of different types of model.

For stationary Box-Jenkins models it was shown that a moving average representation can always be computed. However this representation is not completely general. A weakly stationary time-series consists of the sum of a moving average part and a deterministic part. Specifically, Wold's theorem guarantees that if x_t is weakly stationary, then it can be written as the sum of two components

$$x_t = x_{1t} + x_{2t} \tag{1.1}$$

where x_{1t} is deterministic, in the sense that it may be forecast linearly from previous values with zero mean square error, and

$$x_{2t} = \sum_{j=0}^{\infty} b_j \eta_{t-j}$$

where η_t is a zero mean uncorrelated sequence of random variables.

The deterministic part is also known as the singular component. Its appearance in a weakly stationary process may seem puzzling because it seems to imply that the level of the series depends on time. This is indeed the case for fixed ω. However, the definition of mean-stationarity refers only to an ω average. While the presence of a non-zero singular part generally implies that the time-series is not ergodic, ω averages may still remain stationary. In fact it is possible to construct non-ergodic stationary time-series whose singular parts may be arbitrary finite fourier series.

To show this it suffices to construct a stationary time-series which has singular part $A\cos(\pi bt) + B\sin(\pi bt)$ because any more complicated one may be constructed by summing independent series of this form. Let C_1 and C_2 be independent normal random variables with variance σ^2 with mean zero. Consider the time-series

$$X_t = C_1 \cos(\pi bt) + C_2 \sin(\pi bt)$$

for all t, only C_i being random

Obviously this series has mean zero. Computing $E[X_t X_{t-k}]$ we

find the latter (1.2)

$$= \sigma^2 [\ \sin(\pi bt)\ \sin(\pi b(t-k)) + \cos(\pi bt\ )\ \cos(\pi b(t-k))]$$

$$= \cos(\ \pi bk)$$

so that the process has stationary covariances. Since the magnitudes of the covariances computed using a time average depend on ω in this example, this process is not ergodic. Since for each realization this time-series is periodic it could not be discriminated from a non-stationary one in practice. For this reason it is usually assumed that the overall stochastic process being sampled has no singular component.

This does not mean that in practice time-series do not have singular parts. It is important to point out that Wold's theorem is a theorem about stochastic processes, and not about time-series. Specifically, the process constructed by starting a discrete time-series in its stationary distribution, is different from the process obtained when the initial value is fixed. This distinction is usually not made in discussing time-series but it can make the difference between a non-zero singular component and none.

A time-series conditioned on a fixed initial value usually contains a singular component which must die out more or less rapidly as the stationary solution is approached . For example, in the linear case the solution may contain a sine-wave which damps more slowly as a root of the equation approaches a point on the unit circle . Thus, as the root approaches a point on

the unit circle, the time-series looks more and more singular (non-stationary). However, no matter how slowly the singular part may die out, if it tends to zero then the recursion admits stationary solutions, and the series is stationary in principle. In speaking of time-series then, 'near non-stationarity' means that a slowly dying singular component may appear in the solution of the defining equation, for fixed initial values. For non-linear time-series the 'almost non-stationary ' case seems the most interesting.

The generic non-linear time-series is different in character for the following reason. To construct a non-stationary solution in the linear case, one need only find a non-zero deterministic solution for the difference equation

$$\Theta(X_t) = 0 \tag{1.3}$$

which specifies the time-series. The non-stationary solution is constructed by adding the latter to any stationary solution and applying linearity. The result is truely non-stationary . In the non-linear case this construction generally fails. It seems reasonable to speculate that most (in some sense) useful non-linear recursions will admit true non-stationary solutions only with zero probability. For input errors will interact with a deterministic solution and cause the system to drift away from it.

In fact it is possible to give sufficient conditions for stationarity and ergodicity of a non-linear time-series via a result of Tweedie(1975). The relevant result is as follows.

Let $\{X_n\}$ be a Markov chain taking values in a normed space S (such as R^n) with temporally homogeneous transition probabilities

$$P^n(x,A) = P[X_n \in A \mid X_o = x] \tag{1.4}$$

where $x \in S$, and A is a Borel set. Assume that there is a σ-finite measure ϕ such that that whenever $\phi(A)>0$

$$\sum_n 2^{-n} P^n(x,A) > 0 \quad \text{for every } x \in S \tag{1.5}$$

That is almost all of S is reachable from any $x \in S$. Further assume that for every Borel set A of S, P(x,A) is a continuous function of x in the topology of the norm. The result gives sufficient conditions for the existence of a finite invariant measure μ, that is a measure such that for each Borel set A

$$\mu(A) = \int_S \mu(dy)P(y,A) \tag{1.6}$$

Specifically, (i) μ with this property exists if

$$E[||X_{n+1}|| \mid X_n = x] \text{ is bounded for all } x \tag{1.7}$$

and

(ii) there exists a compact set K with $\mu(K)>0$ and $\epsilon>0$ such that

$$E\{||X_{n+1}|| - ||X_n|| \mid X_n = x\} < -\epsilon \tag{1.8}$$

whenever $x \notin K$.

Whenever such a measure μ exists it can be shown (Tweedie ,1974) that except for $y \in N$ where N is a null set of S (ie. $\phi(N) = 0$)

$$(1.9) \qquad \frac{1}{n}\sum_{i=1}^{n} P^{i}(y,A) \to \mu(A) \text{ as } n\to\infty$$

So that the Markov process settles down eventually into its invariant distribution. This is equivalent to our definition of stationarity.

Most time-series are not Markov because X_n depends on X_{n-1} ... X_{n-k}. Moreover, time-series do not always possess temporally homogenous transition probabilities, as was shown by the example of the second order autoregression with roots on the unit circle in section 1.4 . However, when the time-series defined by a non-linear recursion has temporally homogeneous transition probabilities it is possible to apply Tweedie's result. Consider an autoregression satisfying

$$(1.10) \qquad X_{n+1} = F(X_{n-1},\ldots,X_{n-k}) + \epsilon_n$$

where the error series ϵ_n is normal. If we define

$$(1.11) \qquad Y_n^{p} = [X_n,\ldots,X_{n-p}]$$

then the processes $\{Y_{np}^{p-1}\}_{n=1}^{\infty}$ are Markov whenever $p > k$. Let $||\ .\ ||_p$ be defined by the summed absolute values of the p-tuple and let $S = R^p$. We claim that the following conditions are sufficient for stationarity.

(i) The transition probabilities of the time-series are temporally homogeneous.

(ii) the conditional expectations

$$E(|X_{n+1}||Y_n^{k-1}) \tag{1.12}$$

are bounded on R^k.

(iii) There exists a ball B of radius R such that

$$E[|X_{n+1}||Y_n^{k-1}] \} < R/k \text{ outside } B \tag{1.13}$$

As $X_{n+1} = F(Y_n^{k-1}) + \epsilon_t$, the condition (1.13) depends on the distribution of the errors as well as on F.

We prove this result by first verifying Tweedie's condition for the process $\{Y_{nk}^{k-1}\}_n$ under these assumptions. Without loss of generality we may take R such that

$$E\{|X_{n+1}||Y_{nk}^{k-1}\} < R/k \text{ for all } Y_{nk}^{k-1}. \tag{1.14}$$

To do this take $R' = \text{Max}(R, k \sup E\{|X_{n+1}||Y_{nk}^{k-1}\})$ using boundedness of the conditional expectation. Finally, define

$$\Delta_{nk} = ||Y_{nk}^{k-1}||.$$

We now show that for each p, $\Delta_{nk} > R$ implies $E[|X_{nk+p}||Y_{nk}^{k-1}] < R$. Using the definition of the norm, this

will establish Tweedie's condition for the process $\{Y_{nk}^{k-1}\}$ since $Y_{(n+1)k}^{k-1}$ consists of k coordinates of this form.

To begin observe that if Δ_{nk} exceeds R then $E[X_{kn+1} | Y_{nk}^{k-1}] < R/k$ by (1.13). Note also that (1.13) means that for each $p=1,k-1$, $\Delta_{kn+p-1} > R$ implies

$$E[|X_{kn+p} - Y_{kn+p-1}^{k-1}|] < R/k . \tag{1.15}$$

Using the Markov property, for $p \geq 1$ we can write

$$E[|X_{kn+p+1} - Y_{nk}^{k-1}|] \tag{1.16}$$

$$= E[\, E[|X_{kn+p+1} - Y_{kn+p}^{k-1}|]\, |Y_{nk}^{k-1}]$$

Applying (1.13) to the inner expectation we see that it is always strictly less than R/k as required.

We now argue that because the transition probabilities are not time-dependent the stationary distribution we obtain does not depend on how the X_n were grouped to produce a Markov process and thus that there is a stationary distribution for X_n.

While such series are stationary, there are many examples of stable non-linear recursions with complicated slowly dying solutions. Because non-linear recursions admit a rich variety

of singular parts in the associated time-series, non-linear autoregressive models are appropriate when knowledge of a complicated singular component exists.

5.2 A NONLINEAR AUTOREGRESSIVE/MOVING AVERAGE DUALITY

It will be shown that linear time-series models compare well with non-linear models where series of moving average type are concerned. However it is difficult to make precise the claim that non-linear autoregressive models are generically more singular than linear ones (and thus more useful in modelling singular parts). The computation of singular and moving average parts is not easy in the general case and no procedure seems to be known. However, a type of duality between series of input-output type and series of autoregressive type may be demonstrated formally for certain non-linear processes. This is acheived by a novel expansion technique due to Jones(1978) who employed it to do moment calculations. The result suggests that large non-linearities correspond to large singular parts. Note however that the computed processes are not, strictly speaking, moving averages.

Suppose that a non-linear autoregression satisfies the equation

$$X_{n+1} = \lambda(X_n) + e_t \tag{2.1}$$

where λ is infinitely differentiable. For example $\lambda(x) = \exp(-\frac{1}{2}x^2)$. Consider the family of processes indexed by ξ which

satisfy equations of the form

$$W_{n+1}(\xi) = \xi\,\lambda(W_n) + e_n \tag{2.2}$$

Suppose for now that when $t=0$, $W_t = w_o$ for each ξ. but that each process has a stationary distribution (as a Markov process). Then for each ξ by substitution

$$W_1 = e_1 + \xi\,\lambda(w_o) \tag{2.3}$$

$$W_2 = e_2 + \xi\,\lambda(e_1 + \xi\,\lambda(w_o))$$

$$= e_2 + \xi\,\lambda(e_1) + \xi^2\,\lambda(w_o)\,\lambda^{(1)}(e_1) + \ldots.$$

On expanding λ whenever necessary obtain (2.4)

$$W_n(\xi) = e_n + \xi\,\lambda(e_{n-1}) + \xi^2\,\lambda(e_{n-2})\,\lambda^{(1)}(e_{n-1}) + \text{higher terms}$$

To approximate the original series set $\xi = 1$.

Since the processes have a stationary distribution, we expect that the effect of the starting value w_o will die out. Thus the influence of e_{n-j} must diminish as j grows, and the values of e_{n-j} for small j must have greater influence. Such heuristic reasoning suggests that the expansion obtained above may be a valid decomposition of X_n into stationary processes which are explicit functions of the e_n. That is, a

representation in input-output form. While this is not a true moving average, examination of this result seems to indicate that in this type of representation is more useful when ξ (which determines the size of the singular part) is small. This suggests that a generic non-linear autoregressive time-series will have a slowly dying singular part and will not be usefully expressible in input-output form. In the linear case a time-series with a singular component (non-stationary model) similarly has no moving average representation.

5.3 PREDICTION THEORY FOR MOVING AVERAGES

Many theoretical results exist for the moving average part of a weakly stationary process. See for example Anderson (1971) for general linear theory and Schetzen (1980) for an exposition of some general non-linear theory. Such results seem to show that the emphasis on linear models is often entirely proper. In particular, certain results in prediction theory show that the improvement in prediction attainable through non-linear predictive schemes is limited.

Specifically, when the constituent variables η_t of a linear moving average are independent identically distributed (iid) random variables, classical prediction theory gives the error of the best linear forecast (eg. Deutsch,1965) as follows. Suppose X_t has such a representation

$$X_n = \Sigma_j a_j Y_{n-j} \tag{3.1}$$

Now define

$$\Phi(s) = \left| \sum_{j=-\infty}^{+\infty} a_j e^{ijs} \right| \tag{3.2}$$

and set

$$\Delta^2 = \exp\left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \log \Phi(s) ds \right) \tag{3.3}$$

If $E(Y_j)=0$ and $Var(Y_j)=\sigma^2 < \infty$ then $\Delta^2 \sigma^2$ is the mean square error of optimal one step prediction, that is

$$\Delta^2 \sigma^2 = \inf E\left(\left(X_{n+1} - \sum_{j=0}^{r} b_j X_{n-j} \right)^2 \right) \tag{3.4}$$

where the inf is taken over all finite sequences b_i of real numbers.

It is well known that for Gaussian time-series the linear forecast (obtained by a Box-Jenkins model) is the optimal one in mean square. This is not true for the Non-Gaussian case. Therefore some gain in forecasting power may conceivably be obtained using some non-linear function of preceding observations. However, Kanter(1979) has proved the following result for linear moving averages which assumes only some mild conditions on the distribution of Y_j (which must be iid). Let Ω denote the set of all Borel measurable functions f from R^{∞} into R. Then, letting

$$X_m^n = (\ldots, X_{n-1}, X_n) \tag{3.5}$$

we have

$$(3.6) \qquad \inf_{f \in \Omega} E((X_{n+1} - f(X_n^n))^2) \geq Q(Y_o) \Delta^2$$

where the constant $Q(Y_0)$ is defined to be variance of the Gaussian random variable whose entropy equals the entropy of Y_0. The entropy of a random variable Y is simply $-E\{ \log f(Y) \}$ where f is the density of Y. It is thus just the information function defined earlier evaluated at f(.). For Gaussian sequences $Q(Y_o)$ is just σ^2. The result means that for sequences Y_j with $Q(Y_o) > 0$ and $\sigma^2 < \infty$ then non-linear prediction can improve prediction by at most a factor $Q(Y_o)/\sigma^2$. In particular, if perfect non-linear prediction is possible then perfect linear prediction is possible for moving average processes with finite variance and $Q(Y_o) > 0$.

Thus, the use of Box-Jenkins models to compute predictions for weakly stationary processes with iid moving average representations is sensible provided the random variables in the moving average are not very far from Gaussian. Since however, the random variables of Wold's theorem may not be iid, in principle there may still be some value in other non-linear representations for weakly stationary processes with small singular parts.

For example, Granger and Newbold(1976) considered series of the form

(3.7) $$y_t = g(x_t)$$

where y_t is an instantaneous transform of x_t. If a linear model is identified for y_t then this implies a non-linear model for x_t. Let the inverse transformation of g be denoted by h. If h is known and is in some sense well behaved then an analytical expression for the optimal non-linear h-step forecast $f^x_{n,h}$, of x_{n+h} can be obtained. It may be shown that if the forecastability of the series is measured by

(3.8) $$R^2_{h,x} = 1 - \text{var}(f^x_{n,h}) / \text{var}(x_{n+h})$$

then $R^2_{h,y} > R^2_{h,x}$. That is, the non-Gaussian series constructed in this manner is always less forecastable than the 'original' Gaussian series. Since in addition the construction of non-linear forecasts is more difficult than the construction of linear forecasts, this observation ought to motivate the search for transformations assuming such non-Gaussian series occur in practice. Unfortunately this search is not easy. It must be born in mind that for a series to be Gaussian, all the joint distributions must be multivariate normal, an extremely stringent condition. The most well known set of candidate

transformations is the Box-Cox (1964) class which are of the form

$$y_t' = \{ (x_t + m)^{\theta} - 1 \} / \theta \qquad (3.9)$$

and include logarithmic (θ=0) and linear transformations (θ=1) as special cases. Nelson (1976) found however,that when this transformation was applied to a number of economic time-series the resulting optimal forecasts were often no better than the linear ones, suggesting that in fact the transformations did not give Gaussian series. Thus, without a priori knowledge the search for such a transformation seems problematic.

Other proposals for non-singular non-linear models eg.Nelson&Van Ness(1973) are less well known and difficult to interpret. Because the potential improvement in prediction is limited and identifying non-linear models is difficult, attempts to formulate general non-linear models which are based upon modelling the moving average portion of a stationary time-series will not likely produce great rewards.

5.4 BILINEAR MODELS

Despite the preceding remarks it may sometimes be worthwhile to develop non-linear representations for processes with small singular parts. For one thing, even small increments in predictive power may be highly desirable. In addition, the fitting of such models may be useful as a preliminary procedure in a more detailled structural model fitting protocol .

The most promising class of non-linear weakly stationary non-singular models to appear are the bilinear models of Granger& Andersen(1978) and Subba Rao (1980). While experience with these models is still lacking, it appears that these models are sucessful in formulating non-linear forecasts which acheive some of the gain in predictive power available without resort to structural information. Moreover consideration of these models may lead to useful tools for the empirical detection of non-linearity. These models may be derived as follows. First observe that if a time-series follows the non-linear difference equation

$$X_t = g(X_{t-1}) + \epsilon_t \tag{4.1}$$

then one can derive an AR(1) model by approximating g using the first term of a Taylor expansion. Suppose now that we iterate once before doing the expansion. It is found (again taking first terms only) that

$$X_{t+1} = g(g(X_{t-1}) + \epsilon_{t-1}) + \epsilon_t \tag{4.2}$$

$$= F(X_{t-1}, \epsilon_{t-1}) + \epsilon_t$$

which may be approximated by the model

$$X_{t+1} = aX_{t-1} + bX_{t-1}\epsilon_{t-1} + d\epsilon_{t-1} + \epsilon_t \tag{4.3}$$

assuming g(0)=0, and is differentiable.

This is a bilinear model, so called because X_t is linear both in X_{t-1} and ϵ_t. Under some circumstances it has a useful

structural interpretation. The terms in the expansion may be thought of as representing respectively, the dependence of X_t on X_{t-1}, the interaction of X_{t-1} and ϵ_{t-1}, and the dependence of X_t on an input disturbance ϵ_t. Such a model might also arise if (say) an AR(1) process were observed with multiplicative white noise observation error. For if

$$X_t = \beta X_{t-1} + \epsilon_t \tag{4.4}$$

and

$$Y_t = X_t (1 + u_t)$$

where u_t is a white noise.

then

$$Y_t = \beta Y_{t-1} ((1 + u_t)/(1 + u_{t-1})) + \epsilon_t (1 + u_t) \tag{4.5}$$

which is approximately

$$Y_t = \beta Y_{t-1} (1 - u_{t-1} + u_t) + \eta_t \tag{4.6}$$

These models may be generalized to (4.7)

$$X_t = \sum_{j=1}^{p} a_j X_{t-j} + \sum_{i=0}^{q} b_i \epsilon_{t-i} + \sum_{k=0}^{Q} \sum_{l=1}^{P} \beta_{kl} \epsilon_{t-k} X_{t-l}$$

(note the absence of a term in $X_t \epsilon_t$). This is the model which would arise on approximating

$$X_t = F(X_{t-1}, X_{t-2}, \ldots, X_{t-n}; e_{t-1}, e_{t-2}, \ldots, e_{t-n}) \tag{4.8}$$

by the first terms of the Taylor series for F, where F is

derived by iterating once in an $n-1^{st}$ order non-linear AR model. To compute effectively with this model is it necessary to assume that the errors e_t are independent random variables, in contrast to the linear theory, in which errors need only be uncorrelated. theory, in which errors need only be uncorrelated.

Although there is a large literature on engineering and control applications of such models in which the error series is replaced by a control variable (Bruni,Dipillo&Koch,1974). a complete theory does not exist. For example necessary and sufficient conditions for stationarity and invertibility are not known. Therefore we confine ourselves to a subclass of models (the so-called diagonal models) in the sequel, and actually discuss only one example. This class of models is known to generate useful weakly stationary processes. However, a true bilinear model may be discriminated from a Box-Jenkins model (with iid errors) by examining the covariance structure of the squared series X_t^2. Following Granger&Anderson (1978) we demonstrate these assertions for the particular diagonal bilinear model

$$X_t = \beta X_{t-1} \epsilon_{t-1} + \epsilon_t \tag{4.9}$$

Define $\lambda = \beta\sigma_\epsilon$. Using the standard operator techniques it is found

$$X_t = \epsilon_t + \sum_{k=1}^{\infty} (\epsilon_{t-1} \beta B)^k \epsilon_t \tag{4.10}$$

Multiplying by ϵ_t and taking expectations gives

$$E[X_t \epsilon_t] = \sigma_\epsilon^2 \tag{4.11}$$

Thus the solution given makes the latter product mean-stationary. Returning to the defining relation this implies

$$E[X_t] = \beta\sigma_\epsilon^2 = \lambda\sigma_\epsilon \tag{4.12}$$

To obtain the variance of X_t first square the defining relation to obtain

$$X_t^2 = \beta^2 X_{t-1}^2 \epsilon_{t-1}^2 + 2\beta X_{t-1} \epsilon_{t-1} \epsilon_t + \epsilon_t^2 \tag{4.13}$$

Now take expectations

$$E[X_t^2] = \beta^2 E[X_{t-1}^2 \epsilon_{t-1}^2] + \sigma_\epsilon^2 \tag{4.14}$$

To compute the expectation on the right hand side of the latter equation,use the squared defining relation, multiply by ϵ_t^2 and take expectations. From the independence of the error series we thus get

$$\begin{aligned} E[X_t^2 \epsilon_t^2] &= \beta^2 E[\epsilon_t^2] E[X_{t-1}^2 \epsilon_{t-1}^2] + E[\epsilon_t^4] \\ &= \lambda^2 E[X_{t-1}^2 \epsilon_{t-1}^2] + \sigma_\epsilon^{(4)} \end{aligned} \tag{4.15}$$

This difference equation has a positive solution for

$$E[X_t^2 \epsilon_t^2] \tag{4.16}$$

independent of t iff $|\lambda| < 1$. Thus this model is variance stationary iff $|\lambda| < 1$. To obtain the first order autocovariance ρ_1 procede as follows. Using independence note that

$$(4.17)\quad E[\, X_t X_{t-1} \,] = E[\, (\beta X_{t-1} \epsilon_{t-1} + \epsilon_t)\, X_{t-1} \,]$$

$$= \beta\, E[\, X_{t-1}^2 \epsilon_{t-1} \,]$$

Now substitute for X_{t-1}^2 using the difference equation and take expectations to obtain

$$(4.18)\quad E[\, X_t X_{t-1} \,] = 2\beta^2 E[\, \epsilon_{t-1}^2 \,] E[\, X_{t-2} \epsilon_{t-2} \,]$$

$$= \beta^2 \sigma_\epsilon^4$$

$$= 2(E[\, X_t \,])^2$$

Since (4.19)

$$\rho_1 = \mathrm{cov}(X_t, X_{t-1}) / \mathrm{var}(X_t)$$

a formula for ρ_1 may be computed.

For Gaussian errors, the formula can be made a little more explicit . In particular,

$$\sigma_\epsilon^{(4)} = 3\sigma_\epsilon^4$$

so that

$$E[X_t^2] = \sigma_\epsilon^2 \frac{2\lambda + 1}{1 - \lambda^2} \tag{4.19}$$

Thus, recalling the definition of λ

$$\text{var } X_t = E[X_t^2] - (E[X_t])^2 \tag{4.20}$$

$$= \sigma_\epsilon^2 \frac{(1+\lambda^2+\lambda^4)}{1-\lambda^2}$$

Since the expression for $\text{cov}[X_t, X_{t-1}]$ reduces to $\sigma_\epsilon^2 \lambda^2$ we can compute

$$\rho_1 = \frac{\lambda^2(1-\lambda^2)}{1+\lambda^2+\lambda^4} \tag{4.21}$$

Using this expression it is easily calculated that ρ_1 attains a maximum of .1547 at approximately $\lambda = \pm 0.605$.

Now consider

$$E[X_t X_{t-k}] = \beta E[X_{t-1} \epsilon_{t-1} X_{t-k}] \qquad k>1 \tag{4.22}$$

On substituting for X_{t-1} obtain

$$E[X_t X_{t-k}] = \beta^2 E[X_{t-2} \epsilon_{t-2} X_{t-k} \epsilon_{t-1}] + \beta E[\epsilon_{t-1} X_{t-k}] \tag{4.2}$$

$$= 0 + \beta\sigma_\epsilon^2 E[X_t]$$

$$= (E[X_t])^2$$

Hence

$$\text{(4.24)} \qquad \operatorname{cov}(X_t, X_{t-k}) = E[X_t X_{t-k}] - (E[X_t])^2$$

$$= 0 \qquad k>1$$

Thus the correlation structure of this model is that of an MA(1) Box-Jenkins model.

Consider now the series $Z_t = Y_t^2$ where Y_t is generated by an MA(1) Box-Jenkins model with iid errors e_t then

$$\text{(4.25)} \qquad E[Z_t Z_{t-k}] =$$

$$E[\{a_0 e_n + a_1 e_{n-1}\}^2 \{a_0 e_{n-k} + a_1 e_{n-k-1}\}^2]$$

$$= 0 \qquad \text{for } k>1$$

Thus Z_t again has the correlation structure of an MA(1) series.

In contrast, if $W_t = X_t^2$ where X_t is generated by the bilinear model then it may be shown, by a series of manipulations similar to the forgoing that if the 4th moment of the error series exists

$$\text{(4.26)} \qquad \rho^w_\tau = \lambda^2 \rho^w_{\tau-1} \qquad \text{for } \tau>1$$

while ρ^w_1 is a complicated function of λ. Thus the correlation structure of the squared series is that of an ARMA(1,1) process. In this manner, the linear and bilinear processes may be

distinguished.

It is possible to compare in principle the performances of predictors based on the two models if the process is actually bilinear. If a linear moving average model is used and the covariance structure is known then b in the model

$$X_t = b\,\theta_{t-1} + \theta_t$$

where $\{\theta_t\}$ is white noise, may be calculated as the solution of

(4.27) $$\rho_1 = \frac{b}{1+b^2}$$

$|b|<1$ for invertibility

Then as shown earlier

(4.28) $$\text{var}\,\theta_t = \text{var}(X_t) \,/\, (1 + b^2)$$

Recall that variances and covariances were calculated as functions of λ and σ_ϵ^2. Thus, the relative performance of the two models can be summarized by the ratio

(4.29) $$P(\lambda,\sigma_\epsilon) = \text{var}(\theta_t) \,/\, \text{var}(\epsilon_t)$$

For $\sigma_\epsilon^2 = 1$, P increases from 1.000 to 4.20 as as λ increases from .01 to .75 and P increases from 4.20 to 27.838 as λ increases from .75 to .95 . Thus, for larger values of λ impressive gains in forecasting power are possible. We note that for $|\lambda|>1$ the model becomes non-stationary.

However, whether such gains are commonly realized in

practice depends on the frequency with which bilinear models occur in nature. As Tong&Lim(1980) remark bilinear models do not seem to simulate a large fraction of the types of non-linearity possible, thus the specification of bilinearity is relatively restrictive. Bilinear series do not have interesting singular parts. However, the distributions of the random variables of a bilinear process differ from those of a linear process. For example, Granger&Anderson show the diagonal processes have non-zero third central moments in general , so that the distributions may be skewed, and some higher moments need not exist, assuming the input errors are normal. These observations suggest that bilinear models are best suited to the simulation of time-series which are non-Gaussian moving averages.

Bilinear models have been fitted to the classical data-series such as the Wolfer sunspot numbers and Canadian Lynx data as well as other sample series. It is difficult to compare the results between investigators (e.g. Subba-Rao (1979),Granger&Anderson (1978),Tong&Lim (1980)) owing to the tendency for each to employ individual fitting procedures and to fit their models to different intervals of the series concerned. In addition the performance of these models is typically assessed by omitting a few (usually 10-20) points from the end of the data series and examining the ability of the model to predict these omitted points. While this method cannot be criticized in view of the shortness of these data series, it

means that no one instance is particularly conclusive. Invariably however, an improvement in power of prediction over the linear model between 5 and 25% of mean square error is reported. This suggests that bilinear modelling procedures do provide non-linear predictive power in the absence of structural data.

VI. APPROPRIATENESS OF FULLY NON-LINEAR MODELS: PRACTICE

6.1 VIOLATIONS OF LINEARITY

We have indicated that small non-linearities can be modelled via a general model such as the bilinear class. When models with large non-linearities are admitted potentially greater rewards are possible. However, the richness of structure which such models admit can lead to difficulties in model selection. We have maintained that it is necessary to use structural information in this case. In the following section we discuss the application of this suggestion in practice.

It seems best to think of serious non-linearity as one does non-stationarity - as an assumption forced upon the analyst either a priori, or by the failure of the data to exhibit the pleasant features of linear model data. We will discuss examples. A non-stationary Box-Jenkins model, as discussed earlier is of the form

$$P(B)D^{d}X_{t} = e_{t} \tag{1.1}$$

where D is the difference operator. As noted, this model form is designed to simulate data which is homogeneous in time except for a deterministic or stochastic polynomial trend. A non-linear model, in contrast, must simulate significant deterministic or stochastic inhomogeneities over space, where by space we mean the values taken on by X_t. The most promising classes of models for this purpose in our view are the so called

piecewise linear autoregressive models.

Non-linear models are required when it is known that a linear model leads to a failure to simulate behavior of the system under study. To clarify this criterion, note that linear structure leads to a number of properties which might loosely be called independence properties, because their statements assert that some output property is independent of some input property. Just as different types of non-stationarity justify different choices of differencing order d, violations of different independence properties will indicate that certain non-linear models are appropriate.

We now recall some independence properties of linear models. Most importantly, a linear system is amplitude independent in that if c is a constant

$$c\,L(e_t) = L(c\,e_t) \tag{1.2}$$

That is, inputs of differing amplitudes are output without distortion. This assumption is an unrealistic one for many models since most physical, biological and economic systems exhibit some form of saturation, as well as a linear range. The simplest conceptual example (not the simplest mathematical example) is to define the operator F such that

$$F(e_t)(s) = e_s \tag{1.3}$$

provided $|e_t| < c$ for all $t < s$ and

$$F(e_t)(s) = 0 \qquad \text{otherwise}$$

This is the behavior exhibited by a fuse. Models of flooding and river overflow (Sugawara,1962) as well as earthquakes and stock market prices are natural examples of such phenomena.

Tong&Lim (1980) propose, and fit piecewise linear models which are similar to the fuse in that they contain amplitude dependent thresholds, though they do not have infinite memory. By experimenting with parameters in such models it is possible to illustrate several types of violation of the independence properties possessed by linear systems.

Other manifestations of non-linearity arise from interpreting an amplitude dependence in the frequency domain. Such interpretation is sensible structurally because many natural phenomena are known to be periodic (often because they satisfy linear differential equations) and thus it can happen that the input to a system is a superposition of a finite number of frequency components. When response depends on input amplitude, clear empirical relations may appear in the data between frequency and amplitude. In particular the data contains epochs in which large vibrations of high (low) frequency are interspersed with small vibrations of low (high) frequency. This is observed for example, in records of EEG during sleep. Tong&Lim give the example

$$(1.4)\qquad X_n = 1.6734 - 0.8295X_{n-1} + 0.1309X_{n-2} - 0.0276X_{n-3} + \epsilon_n^{(1)}$$

$$\text{if } X_{n-1} > 0.5$$

$$= 1.2270 + 1.0516X_{n-1} - 0.5901X_{n-2} - 0.2149X_{n-3} + \epsilon_n^{(2)}$$

$$\text{otherwise}$$

where $\text{var}(\epsilon_n^{(i)}) = 0.003^2$, $i=1,2$

which simulates the first mentioned behavior, and another similarly complex example which simulates the opposite case. An important question here seems to be determining the simplest process which gives such behavior.

The superposition principle is an obvious consequence of linearity, which also has consequences in the frequency domain. It may be seen that if a sinusoidal input is applied to a linear system defined by

$$(1.5)\qquad X_n \sum_{j=0}^{\infty} a_{n-j} Y_j$$

then setting

$$Y_j = e^{imj}$$

the output is given by

$$(1.6) \qquad X_n = \sum_{l=0}^{-\infty} a_m e^{im(n-l)}$$

Thus, the output is a sum of attenuated (amplified) and delayed (advanced) sinusoids of the same frequency. This is true of a general linear system, and unsurprisingly may fail for non-linear systems. For example the output of a non-linear system may contain harmonics of the input frequency. Tong&Lim give the example

$$(1.7) \qquad X_n = \begin{cases} -(2+\sqrt{2})\,Y_n - (1+\sqrt{2}) & \text{if } -1 < Y_n \le -1/\sqrt{2} \\ -\sqrt{2}\,Y_n - 1 & \text{if } -1/\sqrt{2} < Y_n \le 0 \\ \sqrt{2}\,Y_n - 1 & \text{if } 0 < Y_n \le 1/\sqrt{2} \\ (2+\sqrt{2})Y_n - (1+\sqrt{2}) & \text{if } 1/\sqrt{2} < Y_n \le 1 \end{cases}$$

which responds to some input sinusoids Y_n with a waveform having frequency double that of the input. This phenomenon is difficult to detect by inspection of the data though it must be born in mind when a system exhibits systematic periodicity of a higher (lower) frequency than can be justified by a linear theory.

From the point of view of time-series analysis the most restrictive feature of linear models is that a stationary linear

system converges to zero in the absence of input, independent of the initial condition. Recall that to see this one puts the system in autoregressive form and a difference equation is solved for the output. It was seen earlier that it is possible to generalize the linear model to simulate processes containing sinusoidal components but that the other types of non-stationarity encompassed by this model lead to processes with unbounded variance and are thus of limited interest.

In contrast non-linear recursions may have deterministic solutions such that for various initial conditions $\{ X_o \in I \}$ there are distinct limit sets A(I) with the property

(1.8) $\quad x \in A$ iff for all $\epsilon > 0$ $|x - X_n| < \epsilon$ infinitely often

We will call such a set an attractor. In particular if the state space is partitioned into a finite number of initial sets I_i such that for each $i>1$ $A(I_i)$ is a single point then we say the system possesses multiple equilibria. If for some initial set I A(I) contains a closed trajectory of the system we say the system admits a cycle. If there exists an I containing an open neighbourhood of A(I) and A(I) contains only finite closed trajectories then the system admits stable limit cycles. Finally, trajectories of non-linear recursions need not be closed even though they are bounded. In this case it is said the system admits chaotic solutions. (Li&Yorke,1975). When limit cycles or chaotic solutions exist deterministic non-linear

phenomena may account for most of the variance of the series, although it may present a random appearance to some conventional analyses (May,1974; Bunow&Weiss,1979).

The paper of Bunow &Weiss illustrates the complexities which ensue when simple deterministic recursions are analysed as linear time-series via their autocovariance functions. These examples are non-linear time-series which are entirely singular. Several recursions were studied including the discrete logistic recursion (LI&Yorke,1975)

$$X_{n+1} = r X_n (1 - X_n) \tag{1.9}$$

and the triangle recurrence (Guckenheimer,Oster&Ipaktchi(1977)

$$X_{n+1} = 1 - 2 \mid X_n - .5 \mid \tag{1.10}$$

by systematically varying parameters and initial values.

For the logistic recursion, the most common case (eg. $r=3.64$, $X_0 =.878$) is an almost uniform comblike autocorrelation function of high amplitude extending to very large lags, in excess of 100 points. Less common (eg. $r=3.75$, $X_0 =.374$) was a rapidly decaying oscillatory autocorrelation function with small spindles at large lags. Sometimes (eg. $r=4.00$, X_0) the autocorrelation function was indistinguishable from that of a sequence of independent random variables, or at the other extreme, was periodic. The triangle recursion graph is similar to that of the logistic recursion with $r=4.00$ so it is perhaps not surprizing that its autocorrelation function was usually

identically zero.

Limit cycles, chaos, and other non-linear phenomena characterize the rich variety of singular parts in non-linear time-series models. This fact is practically important because of examples in which a priori models which predict non-linear phenomena and yet investigators have persistently fitted linear models to data. This is true particularly of the ecological literature (cf. Campbell&Walker,1977; Tong,1977) as well as certain phenomena in economics (cf. Granger&Anderson,1978).

6.2 THE THRESHOLD MODELS OF TONG&LIM

Tong&Lim(1980) proposed a variety of non-linear autoregressive models, with the property that different linear difference equations were satisfied for different values of X_t. As noted, such models can simulate many of the non-linear phenomena described above. However, because this class of models is considerably richer than the linear class there is some difficulty in conducting the identification procedure. The fitting procedure for one type of model is described as follows.

Let $\underline{P}$ be a partition of the real line corresponding to the $l+1$ points $\{r_0, r_1, \ldots r_l\}$. arranged in increasing order. The r_i will be referred to as thresholds. Define $R_j = (r_{i-1}, r_i]$. Then, a self-exciting threshold autoregressive model of order $(l;k,\ldots,k)$ or SETAR$(l;k,\ldots,k)$ (k is repeated l times and gives the order of the linear autoregression for each of l

regimes) has the form

$$X_n = a_o^{(J)} + \sum_{i=1}^{k} a_i^{(J)} X_{n-1} + \epsilon_n^{(J)}, \qquad (2.1)$$

where we write

$$\underline{X}_n = (X_n, X_{n-1}, .., X_{n-k+1})^T$$

and J is a random variable such that $J = j$ if $\underline{X}_{n-1} \in R_j^{(k)}$ for $j=1,l$. It is assumed that the errors are Gaussian (the series is not) independent in different regimes, and that $R_j^{(k)}$ is a cylinder set of R^k, of the form $R \times R \times .. R_j \times ... \times R \times R$ depending on some fixed lag d. Because these models are autoregressions they are always invertible.

To fit such a model, it is necessary to estimate the thresholds, the critical lag d and the respective orders of the linear autoregressive regimes. Tong&Lim describe the procedure followed in fitting a SETAR(2,k_1,k_2). To begin, Q candidate thresholds t_q and D critical lags are selected. In this case each threshold distinguishes two regimes. Method of model comparison used is Akaike's criterion which in this case is proportional to

$$AIC(k) = N \ln(RSS / N) + 2k \qquad (2.2)$$

where N is the number of observations, k the number of fitted parameters and RSS denotes the residual sum of squares. On each

data set determined by each fixed t_q and d the autoregressive orders are determined to minimize AIC(k) for $0 \leq k \leq K$ where K is some fixed maximum order. Because the errors on each regime are independent one may next write

(2.3) $$AIC(t_q) = AIC(\hat{k}_1) + AIC(\hat{k}_2)$$

This criterion is minimized in turn to estimate t_q. The latter procedures establish a minimum AIC model for each value of d.

A difficulty arises in comparing models with different values of d because of the asymptotic nature of Akaike's statistic. In particular, the selection of a different lag results in a different effective number of observations. Tong&Lim apply a normalization and thus compute

(2.4) $$AIC(d) = AIC(\hat{r}_i) / (n - \max(d, K))$$

This procedure produces semi-automatic estimates for all the necessary parameters. As in the linear case residuals may be plotted and analysed and the model possibly rejected.

It is difficult to evaluate the sucess of Tong&Lim's procedures at this relatively early stage. In fitting classical data series such as the Canadian Lynx data and the Wolfer Sun spot numbers they report 'disappointing' results. Nonetheless, for the fitting periods chosen by Tong&Lim, modest improvements in prediction power over the linear (and bilinear) model were reported. For example for the Wolfer sunspot numbers from 1770-1869 the mean square error of one step prediction over the next

20 years was 346.6 for an AR(2) model(Box&Jenkins),

$$(2.5) \qquad X_t = 14.70 + 1.425X_{t-1} - 0.731X_{t-2} + \epsilon_t$$

293.4 for the linear model in combination with a bilinear model fitted by Granger&Anderson(1978) to the residuals ϵ_t

$$(2.6) \qquad \epsilon_t = -.0222\epsilon_{t-2}\eta_{t-1} + 0.202\epsilon_{t-1} + \eta_t$$

and 267.6 for the SETAR(2;3,4) model fitted by Tong&Lim

$$(2.7) \quad X_n = 5.269+1.889X_{n-1} - 1.5289X_{n-2} + 0.3039X_{n-3} + 0.3387X_{n-4} + \epsilon_n^1$$

$$\text{if } X_{n-3} < 36.6$$

$$= 0.3900+1.1366X_{n-1} - 0.3645X_{n-2} + 0.0524X_{n-3} + \epsilon_n^2$$

$$\text{if } X_{n-3} > 36.6$$

This seems an impressive gain in predictive power, however, there appears to be some non-stationarity in this data which is difficult to model. One step predictions over a longer period deteriorate. The predictions of a SETAR(2;4,2) fitted to the 1837 to 1920 data deteriorate after 1944.

For the (logarithmically transformed) Lynx data (1821-1920) a linear AR(12) model, selected using Akaike's criterion Tong(1977) had a mean square error of prediction of .018 , while a SETAR(2; 8,3) had mean square error .0144 and a bilinear model

fitted by Subba-Rao (AR order 11 with cross terms up to $X_{t-8}\,\epsilon_{t-10}$) had the same mean square error as the linear model over 14 subsequent years. The most striking feature of these results is that the mean-square errors are so similar. Certainly no clear margin of superiority is evident for piecewise linear over bilinear models, or even over linear ones. It seems consistent with the evidence to speculate that there is a practical limit on the expected improvement over a linear model without resort to structural data, and both bilinear and piecewise linear models are just rich enough to be close to the optimum in these cases.

In contrast a fitted model for flow of the Kanna river in Japan as a function of rainfall and past flow gave a reduction of 18% in mean square error against an (unspecified) competing linear model over 86 time points. This is their most clearly sucessful model in this respect. It is a complicated model. There is one (rainfall) threshold distinguishing two regimes. One involves log riverflow X_t to lag 5 and rainfall Y_t to lag 4. The other regime involves each variable to lag 2. While the model may be criticized for its complexity, before presenting the fitted model Tong&Lim note that seasonal variations of Japanese rivers are quite regular due to the rather well defined rainy season there. In addition, the ground soil is rarely dry so that variations in the water table are small. Since rainfall is not dissipated instantly, it is reasonable that present

riverflow is a function of past river flow. They conclude that it is reasonable to model river flow data as a function of rain fall which may contain thresholds. This model thus has some structural justification, unlike those for the classical series. In view of what we have said so far the success of this model is predictable. Akaike notes in the discussion that in his experience (implementing control strategies for industrial processes) success had been acheived by characterizing the system physically in order to chose the particular non-linearitities to be modelled. He does not suggest that a coherent theory need exist, only that the appropriate conditioning variables should be chosen from physical considerations, or simple examination of the data.

In evaluating Tong&Lim's approach it is useful to focus on Chatfield's questions. 'How can we tell if a given time-series is non-linear?' and 'How can we decide if it is worth trying to fit a non-linear model?'. In our view, a serious attempt to answer these questions even in particular analyses is a large step towards a successful model. If so there would be much value in future research to determine the simplest threshold models which produce particular violations of linearity, and to catalogue examples where threshold models are scientifically justified.

6.3 SUMMARY

The problems inherent in identifying time-series models were discussed and it was shown that structural information can be judiciously utilized when a parametrization is to be selected. Recent proposals for non-linear time-series models were examined in this light. Weakly stationary time-series were seen to consist of deterministic and moving average components. The potential for non-linear modelling of the moving average component was shown to be limited, but it was suggested that worthwhile improvements could result from modelling this component with a weakly non-linear (bilinear) data prescription. Conversely, we suggest that modelling of the deterministic component might best be accomplished via the strongly non-linear (piecewise linear) data prescriptions, but that for this purpose some structural information may be required.

APPENDIX A - SOLVING A HOMOGENEOUS DIFFERENCE EQUATION

Let Z(t) be a real function of an integer time index satisfying

$$(A.1) \quad P(B)Z(t) = Z(t) - p_1 Z(t-1) - p_2 Z(t-2) - \ldots - p_n Z(t-n) = 0$$

We will discuss the method of solution of such a recursion. It turns out that the solutions are completely analogous to the solutions of homogeneous linear differential equations . Thus the solutions can be obtained by first substituting for Z(t) functions of the form

$$(A.2) \quad (i) \quad A r^t \quad \text{or} \quad (ii) \quad A t^m r^t$$

Next, determine r from the resulting algebraic equation, and count linearly independent solutions to see that the general solution has been obtained. Another way to express this procedure is as follows. Note that because P(B) is formally a polynomial in B, it can be factored using the fundamental theorem of algebra as

$$(A.3) \quad (I - r_1 B)(I - r_2 B) \ldots (I - r_n B)$$

where

- I is the identity operator on sequences
- B is the backward shift operator on sequences defined by $B X_t = X_{t-1}$
- r_i are complex scalars

Because IB=BI, each of the bracketed factors commute, any factor we choose may be written last. Also note that any operator polynomial T(B) operating on the zero sequence gives zero by linearity. These facts together reduce the solution of the difference equation to the solution of

$$(A.4) \qquad (I - rB)^m Z_t = 0$$

Thus for each root of multiplicity m there corresponds an arrangement of the factors so that the operator in the latter equation appears last. If this equation is solved, we have a solution of the original equation, for the effect of the last m factors is to annihilate Z(t). To solve this equation use an inductive approach. First note that for n=1 the obvious solution is Ar^t

We claim that the equation (A.4) has m independent solutions of the form

$$(A.5) \qquad A\, t^j\, r^t \qquad j=1,m$$

Since the result is true for m=1 it suffices to show

$$(A.6) \qquad (I - rB)\{ t^{j} r^{t} \} = K_{j-1}\, t^{j-1} r^{t} + K_{j-2}\, t^{j-2} r^{t} + \ldots + K_{0}$$

but

$$(A.7) \qquad (I - rB)\{ t^{j} r^{t} \} = t^{j} r^{t} - r\,(t-1)^{j} r^{t-1}$$

$$= \{ t^{j} - (t-1)^{j} \}\, r^{t}$$

So that the degree of the putative solution is reduced by 1 by the operator I-rB. Thus, the result follows because applying each of the m factors one after another must eventually annihilate the original form (A.5).

BIBLIOGRAPHY

AKAIKE,H. (1969) Fitting autoregressions for prediction. Annals of the Institute of Statistical Mathematics , 21 ,No.2,243-247.

AKAIKE,H. (1973) Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory ,(B.N.Petrov&F.Csaki,ed),pp.267-281,Budapest Akademai Kiado.

ANDERSON,T.W. (1971) The Statistical Analysis of Time-Series ,New York, Wiley.

BORA-SENTA,E.&KOUNIAS,S.(1979) Parametric estimation and order determination of autoregressive models. In Analysing Time-Series ,(O.D. Anderson, ed.),pp 93-103, Amsterdam, North Holland.

BOX,G.E.P.& COX,D.R. (1964) An analysis of transformations. Journal of the Royal Statistical Society, Series B , 26 , 211-252.

BOX,G.E.P.&JENKINS,G.M. (1970) Time-Series Analysis, Forecasting and Control ,San Francisco, Holden Day.

BRUNI,C., DIPILLO,G.& KOCH,G.(1974) Bilinear systems: an appealing class of 'nearly linear' systems in theory and applications. IEEE Transactions in Automatic Control , TAC-74 , 334-348.

BUNOW,B. &WEISS, G.H. (1979) How chaotic is chaos? Chaotic and other 'noisy' dynamics in the frequency domain. Mathematical Biosciences , 47 ,221-257.

CAMPBELL,M.J. & WALKER, A.M. (1977) A survey of statistical work on the Mackenzie river series of annual Canadian lynx trappings for the years 1821-1934 and a new analysis. Journal of the Royal Statistical Society, Series A , 140 ,411-431.

DEUTSCH, R. (1965) Estimation Theory , Englewood Cliffs, N.J. Prentice Hall.

DURBIN, J. (1960) The fitting of time-series models. Review of the International Institute of Statistics , 28 ,233-289.

FRIEDMAN, J.H. & STUETZLE W. (1981) Projection Pursuit Regression. Journal of the American Statistical Association , 76 ,817-823.

GERSH, W. & SHARPE, D.R. Estimation of power spectra in finite order autoregressive models. IEEE Transactions in Automatic Control , AC-18 , 367-379.

GRANGER, C.W.J. & ANDERSEN, A.P. (1978) Introduction to Bilinear Time-Series Models . Gottingen, Vandenhoeck & Ruprecht.

GRANGER,C.W.J. & MORRIS, M. Time-series modelling and interpretation . Journal of the Royal Statistical Society, Series A , 139 , 246-257.

GRANGER, C.W.J. & NEWBOLD, P. Forecasting transformed series. Journal of the Royal Statistical Society, Series B , 38 , 189-203.

GUCKENHEIER, J., OSTER, G. & IPAKTCHI, A. (1977) Dynamics of density dependent population models, Theoretical Population Biology , 4 , 401-415.

HANNAN, E.J. & QUINN, B.G. (1979) The determination of the order of an autoregression . Journal of the Royal Statistical Society , Series B . 49 , 190-195.

HOSKING, J.R.M. (1978) A Unified derivation of the asymptotic distributions of goodness of fit statistics for autoregressive time-series models. Journal of the Royal Statistical Society, Series B , 40 , 341-349.

JONES, D.A. (1978) Non-linear autoregressive processes. Proceedings of the Royal Society of London, A , 360 ,71-95.

JONES, R.H. (1975) Fitting autoregressions. Journal of the American Statistical Association , 70 , 590-592.

JONES, R.H. (1976) Autoregression order selection. Geophysics , 41 , 771-773.

JONES, R.H. (1981) Fitting a continuous time autoregression to discrete data. In Applied Time-Series Analysis II ,(D.F. Findley, ed.) Academic 1981 , pp 651-674.

KANTER, M. (1979) Lower bounds for non-linear prediction error in moving average processes. Annals of Probability , 7 ,128-138.

KOOPMANS, L.H. (1974) The Spectral Analysis of Time-Series . New York, Academic Press

LI, T.Y. & YORKE, J.A. (1975) Period three implies chaos. American Mathematical Monthly , 82 ,988-992.

LUDWIG, D. (1975) Persistence of dynamical systems under random perturbations. SIAM Review , 17 ,605-640.

MAY, R.M. (1974) Biological populations with non-overlapping generations : stable points, stable cycles & chaos. Science , 186 ,645-647.

NELSON, H. (1976) The Use of Box-Cox Transformations in Economic Time-Series, An Empirical Study , Unpublished Ph.D. thesis, Economics Department, University of Southern California at San Diego.

NELSON, J.Z. & VAN NESS, J.W. (1973) Formulation of a non-linear predictor Technometrics , 15 ,1-12.

OZAKI, T. (1977) On the order determination of autoregressive-moving average models. Applied Statistics , 26 , 290-301.

PARZEN, E. (1974) Some recent advances in time-series modelling. IEEE Transactions in Automatic Control , AC-19 ,713-715.

QUENOUILLE, M.H. (1947) A large sample test for goodness of fit of autoregressive schemes. Journal of the Royal Statistical

Society, Series A , 110 ,123-129.

RAMSEY, F.L. (1974) Characterization of the partial autocorrelation function. The Annals of Statistics , 2 ,1296-1301.

SCHETZEN, M. (1980) The Volterra and Wiener Theories of Nonlinear Systems , New York, Wiley.

SCHWARTZ, G. (1978) Estimating the dimension of a model. Annals of Statistics , 6 ,461-464.

SHIBATA, R. (1976) Selection of the order of an autoregression by Akaike's information criterion. Biometrika , 63 ,117-126.

STONE, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society, Series B , 39 ,44-47.

STONE, C.J. (1977) Consistent non-parametric regression. Annals of Statistics , 5 ,595-645.

SUBBA-RAO, T. (1980) Discussion on paper by Tong&Lim (1980). Journal of the Royal Statistical Society , 42 ,278-280.

SUGAWARA, M. (1962) On the analysis of run-off structure of several Japanese rivers. Japanese Journal of Geophysics , 2 ,1-76.

TONG, H. (1977) Some comments on the Canadian lynx data. Journal of the Royal Statistical Society, Series A , 140 ,432-436, 448-468.

TONG, H. & LIM, K.S. (1980) Threshold autoregression, limit cycles and cyclical data. Journal of the Royal Statistical Society , Series B , 42 ,245-292.

TWEEDIE, R.L. (1974) R-theory for Markov chains on a general state space I: solidarity properties and R-recurrent chains , Annals of Probability , 2 , 840-864

TWEEDIE, R.L. (1975) Sufficient conditions for ergodicity and recurrence of Markov chains on a general state space. Stochastic Processes and their Applications , 3 ,383-403 .

WOLD, H. (1953) A Study in the Analysis of Stationary Time-Series Second edition. , Stockholm, Arquist and Wiksell.

WHITTLE, P. (1952) Tests of fit in time-series. Biometrika , 39 , 309-318.

YULE, G.U. (1927) On the method of investigating periodicities in distributions of series with special reference to Wolfer's sunspot numbers. Philosophical Transactions of the Royal Society of London, Series A , 226 , 267-298.