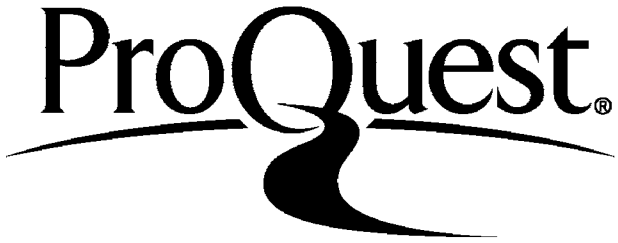
INVESTIGATION OF THE STRUCTURAL PROPERTIES OF KALMAN FILTER MODELS FOR FORECASTING NON-STATIONARY TIME SERIES

by

Janice Margaret Stone
Royal Holloway College

Thesis submitted to the University of London for the degree of Doctor of Philosophy

September 1981



ABSTRACT

This work examines structural properties of dynamic linear models (DLMs) and considers their implications for predictors and minimum-variance estimators of these models. The techniques employed are those of statistical time series analysis and modern control theory, and particular use is made of the concept of observability.

Structural properties are derived and considered for the constant forecast model, the polynomial model of arbitrary degree, both with and without a forward shifted forecast function, and for the more general non-seasonal models with an asymptotic forecast function. Several equivalence theorems are established for these DLMs. It is shown that the invertibility of the time series model is equivalent to the stability of the estimation scheme for the DLM in the equilibrium state. It follows that all predictors of a stable DLM are identical in the steady state to those of the Box-Jenkins forecasting schemes. Furthermore, the observability requirement yields an upper bound on the dimension of the state vector, and a lower bound is necessary to avoid specified restrictions on the equivalences. Examples are considered which show that the practical requirement that the system error covariance matrix be diagonal can further restrict the equivalence.

In addition, the Cramér-Rao bound is considered for estimators of the state vector. It is shown that the information matrix is invertible, and there is a unique estimator which achieves the Cramér-Rao bound if and only if the DLM is observable. This result is also discussed in varying degrees of generality.

TABLE OF CONTENTS

PREFACE

No part of this thesis has been, or is being currently submitted for any degree, diploma or other qualification at any other University.

Unless otherwise stated, all the material of Chapters 6-11 is believed to be original, except for the extra proof of the convergence of A on page 92.

I acknowledge the financial support of an S.R.C. Case Award, jointly with Ferranti Computer Systems Limited, which enabled me to undertake this work.

I am grateful to my Supervisor at Ferranti, Dr. J.R. Moon, for his help and guidance, also to several members of the Statistics and Mathematics Departments of Royal Holloway College for helpful suggestions, especially Professor H.J. Godwin and Dr. D.L. Yates. I would especially like to thank my Supervisor, Mr. E.J. Godolphin; without his help and encouragement, this work would not have been possible.

Many thanks are due to Mrs. B. Rutherford for her excellent typing, and to Miss H. Whittaker for drawing the diagrams.

CHAPTER 1

INTRODUCTION

The analysis of data collected on processes which evolve in time has occupied the attention of scientific workers for several decades. Many of the data series published by Government departments are examples of time series. In general, observations of such time series are dependent, as in population series, where the size of a population in any one year is dependent on population figures in previous years. In many cases, this dependence is due to some underlying process, which may be known to the analyst. For example, the observed position of an aircraft at any time t may be dependent on the position and velocity at previous times through the equations of motion. Similarly, the sales of an item in a particular month may depend on demand in previous months, and on the rate of increase of demand. If the structure of the underlying process is unknown, then a model is fitted from consideration of the data only. If the structure is known, then this can be used in the formulation of a model, and the elements of the underlying process (e.g. velocity, rate of change of demand) can be estimated from the observations. The model can be verified from the data and is then available for the prediction of future observations.

Most time series are not purely deterministic, hence accurate modelling requires the use of random processes. In particular, the classical statistical models for describing time series data are the linear stationary processes which are defined in Chapter 2. For example, Wold (1938) discusses in detail the

autoregressive (AR), moving average (MA) and autoregressive-moving average (ARMA) models. Many naturally occuring time series (e.g. population statistics) are not stationary, but other workers were able to describe these series by suitably transforming the data to yield a stationary series. The autoregressive integrated moving average (ARIMA) model seems to be the most popular model for non-stationary time series (Box and Jenkins 1970). Here the data is rendered stationary by differencing. Other forms of non-parametric transformation are sometimes used, such as the taking of logarithms. Recently, attempts have been made to generalize the ARIMA models to certain non-linear models, for example Tong and Lim (1980) and Priestley (1980).

As a general rule, the statistician has been concerned primarily with the prediction of future observations. The estimation of the underlying process has usually been of secondary importance. For this reason, a typical approach might be to fit the 'best' linear stationary model to the data, transformed if necessary, without special regard for the underlying process. One of the drawbacks of this approach, however, is that the parameters of the model are not always straightforward to interpret. In particular, if there should be any change in the underlying process, it is not necessarily clear how this would affect the parameters of the model.

The use of random processes for the analysis of time series data is also of concern to the control engineer. Traditionally, he exploits the underlying process to obtain the model, which is based on physical considerations and is written in the so-called state-space formulation.

Each component of the state space is intended to have some physical meaning: for example, distance, velocity, acceleration when modelling the movement of an aircraft. Indeed, the control engineer is primarily interested in the estimation and prediction of the state of the underlying process from the observations available. In general, he is less concerned with the prediction of future observations. Several state estimation schemes, usually known as filters, have been proposed. One of the most popular, and one which appears most meaningful is the recursive filter suggested by Kalman and his co-workers (e.g. Kalman, 1960, Kalman and Bucy 1961, Kalman 1963a, Kalman, 1963b) which is now known as the Kalman filter.

In their pioneering work, Harrison and Stevens (1976) proposed using the state space representation together with the Kalman filter for time series analysis in a Bayesian framework. The state vector can be chosen in a way which is meaningful, and thus any changes in the process are more easily incorporated into the model. For example, in modelling sales of a product which is slowly increasing in popularity, it seems sensible to choose a linear growth model, which can be expressed in terms of a level and a slope. These authors call their models 'dynamic linear models' (DLMs) and they assume normality of the noise components. This assumption is not required for the basic results, but it enables distributional results to be obtained, which are helpful in decision making.

The principal objective of the work in this thesis is to investigate the properties of the DLM as a

representation for time series. Most of the work will be confined to univariate time series models. Much of the emphasis will be on the comparison of the predictors of the DLMs with those of the ARIMA models introduced by Box and Jenkins. It should be noted that the ARIMA models have constant parameters, which are estimated from all the available data, while the parameters produced by the Kalman filter algorithm vary as more data become available. This makes it difficult to compare the two approaches. However, it is shown in what follows (Chapter 5) that in some circumstances the parameters of the DLM converge to constants as $t \to \infty$. This is the equilibrium state of the filter, and in these conditions, comparisons of the two approaches can take place.

In Chapter 2, we define a stationary time series, and introduce the most widely used examples, namely the autoregressive, the moving average and the autoregressive-moving average models. The autocovariance function and the spectral density function are defined, and the relationship between the parameters of the model and the autocovariances is illustrated. To satisfy the need for more general models, we introduce the ARIMA model, which is non-stationary. We then discuss some of the methods available for estimating parameters of the models defined, and mention attempts to fit these models to real data.

Chapter 3 discusses methods for forecasting the models defined in Chapter 2.

In Chapter 4, we introduce the state-space representation for linear models. In particular, we define the DLM

introduced by Harrison and Stevens and discuss in what respects it may be more meaningful than the conventional time series models. We also discuss the estimation method defined by the Kalman filter, and consider some of the classes of models which can be expressed as DLMs. We consider the question of forecasting DLMs and discuss some aspects of fitting DLMs to time series data. We conclude this chapter with a comparison given by Akaike (1974a) of the state-space representation and the classical time series model. The state-space representation for linear models has been used by control engineers for some time. In the analysis of the model, we shall draw on some control theory concepts and tools. These are introduced in Chapter 5. The z-transform and the concept of stability play an important part in this respect, as does observability. Jury's criterion for stability (Jury, 1964) is included in Chapter 5.

The constant forecast model is considered in Chapter 6. This, being not only the simplest case, but also one of the most widely used models, is given separate attention. In particular, we discuss the implication of constant forecasts on the structure of the DLM. The Harrison-Stevens (1976) steady model is considered, and it is shown that the predictors of the models are the same as the predictors of a subset of the ARIMA (0,1,1) models. Methods of extending the model to include all ARIMA (0,1,1) models are discussed. It is shown that the dimension of the model needs to be increased for generality. We investigate the effects of assuming that the covariance matrix is diagonal,

and produce a model with a diagonal covariance matrix for which the predictors are identical to those of any ARIMA (0,1,1) model. The effects of a possible restriction on the model are also discussed.

In Chapter 7, we discuss the general polynomial model, where all of the forecasts follow a polynomial of degree $d - 1$. It is shown that if the DLM converges to an equilibrium state, then the predictors of the model are identical to those of the ARIMA (0,d,q) model ($q \leq d$) if and only if the model is stable (or invertible). The intuitively obvious choice for the dimension of this model is d, but we find that this implies that the class of ARIMA (0,d,q) models for which the identity holds is rather restricted. From this, and the concept of observability, we conclude that a model of dimension $d + 1$ is preferred. We also have the result, which is surprising from an intuitive viewpoint, that if the range of the parameters is to be unrestricted, then the model must be singular.

Several of the results of Chapters 6 and 7 are described briefly in the paper by Godolphin and Stone (1980).

These ideas are extended in Chapter 8, where the forward-shifted polynomial model is discussed. Here the first few predictors seem to follow no particular pattern, but all subsequent predictors follow a polynomial path. If the degree of the polynomial is $d - 1$, and the shift is r, then we find that the dimension of the DLM must be $d + r + 1$. This means there is no non-singular model which can possibly describe this situation.

Chapter 9 contains a further extension of these ideas to the general case. We discuss conditions for the

equivalence of the predictors of DLMs and ARIMA (p,d,q) models. It is shown briefly that the results are easily extended to cover seasonal models.

By definition, the Kalman filter provides linear minimum variance estimators of the state variables. If the distribution is normal, then the minimum variance estimators are linear functions of the observations, hence in this case the Kalman filter provides the minimum variance estimator. A useful insight into this topic is obtained by deriving the Cramér-Rao bound for this problem, and where possible, the estimator which attains the bound. This is the subject of Chapter 10. As expected, we find that in the special case of normal distribution and no plant noise, the estimator produced by the Kalman filter is indeed the minimum variance estimator. This is also true for multivariate observations.

CHAPTER 2

TIME SERIES

2.1 Introduction

This chapter is devoted to univariate time series models. It is assumed we have a data series $y_1, y_2, \ldots, y_N$, and we require a model which fits the data so that inferences and predictions can be made. We shall only be concerned with linear models, although recently theories of non-linear time series models have been proposed by several authors, for example, Tong and Lim (1980) and Priestley (1980).

We first define stationary models, and then relax some of the assumptions of stationarity. Most of the models of interest will be non-stationary, but it is usual to consider inference questions for stationary models only.

2.2 Stationary Time Series

A time series $\{y_t\}$ is said to be strictly stationary if $y_{t_1}, y_{t_2}, \ldots y_{t_N}$ has the same joint distribution as $y_{t_1+k}, y_{t_2+k}, \ldots y_{t_N+k}$ for any time points $t_1, \ldots t_N$ and any integer k (positive or negative).

If we also assume that $y_{t_1}, y_{t_2}, \ldots y_{t_N}$ has a finite autocovariance matrix, then the above definition implies that $E(y_t)$ is a constant, and $\text{cov}(y_t, y_{t+h})$ is equal to $\text{cov}(y_s, y_{s+h})$ for all s,t,h and is therefore a function of h only. We adopt this as our definition of a stationary time series. This is sometimes defined as a weakly stationary time series, or a time series

stationary to the second order.

We define the autocovariance function $\{\gamma_h;\ h \text{ integer}\}$ and the autocorrelation function $\{\rho_h;\ h \text{ integer}\}$ for $\{y_t\}$ as follows:

$\gamma_h = \text{cov}(y_t, y_{t+h})$ i.e. γ_h is the autocovariance of $\{y_t\}$ of lag h

$\rho_h = \gamma_h/\gamma_o$ i.e. ρ_h is the autocorrelation of $\{y_t\}$ of lag h.

These moments satisfy

$$\gamma_h = \gamma_{-h},\quad \rho_h = \rho_{-h},\quad \rho_o = 1,\quad |\rho_h| < 1,\ h \neq 0.$$

Three examples of stationary time series are often quoted in the literature.

The moving average model of order q:MA(q) model

The moving average model of order q is defined by

$$y_t = \varepsilon_t + \beta_1\varepsilon_{t-1} + \ldots + \beta_q\varepsilon_{t-q} \tag{2.1}$$

where $\{\varepsilon_t\}$ is a sequence of uncorrelated random variables with a common variance $\text{var}(\varepsilon)$ i.e.

$$E(\varepsilon_t) = 0,\ E(\varepsilon_t\varepsilon_k) = \delta_{t,k}\text{var}(\varepsilon),$$

where $\delta_{t,k}$ is the Kronecker delta. Thus $E(y_t) = 0$ and the autocovariances of $\{y_t\}$ are given by

$$\gamma_o = (1+\beta_1^2+\ldots+\beta_q^2)\text{var}(\varepsilon)$$

$$\gamma_h = (\beta_h+\beta_1\beta_{h+1}+\ldots+\beta_{q-h}\beta_q)\text{var}(\varepsilon) \qquad 1 \leq h \leq q$$

$$\gamma_h = 0 \qquad h > q .$$

The autoregressive model of order p:AR(p) model

The autoregressive model of order p is defined by

$$y_t+\alpha_1 y_{t-1}+\ldots+\alpha_p y_{t-p} = \varepsilon_t \tag{2.2}$$

where $\{\varepsilon_t\}$ is a sequence of uncorrelated random variables defined as above. This model is stationary provided the roots of the polynomial

$$\alpha(z) = 1 + \alpha_1 z + \ldots + \alpha_p z^p \tag{2.3}$$

are greater than one in modulus. The autocorrelations of $\{y_t\}$ satisfy the Yule-Walker equations:

$$\begin{aligned}
\rho_1 + \alpha_1 + \alpha_2\rho_1 + \ldots + \alpha_p\rho_{p-1} &= 0 \\
\rho_2 + \alpha_1\rho_1 + \alpha_2 + \ldots + \alpha_p\rho_{p-2} &= 0 \\
&\vdots \\
\rho_p + \alpha_1\rho_{p-1} + \alpha_2\rho_{p-2} + \ldots + \alpha_p &= 0 \\
\rho_h + \alpha_1\rho_{h-1} + \alpha_2\rho_{h-2} + \ldots + \alpha_p\rho_{h-p} &= 0 \qquad h \geq p.
\end{aligned} \tag{2.4}$$

The autoregressive-moving average model of order (p,q)

This is sometimes referred to as the ARMA (p,q) model and is defined by

$$y_t + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} = \varepsilon_t + \beta_1\varepsilon_{t-1} + \ldots + \beta_q\varepsilon_{t-q}. \tag{2.5}$$

This model is stationary provided $\alpha(z)$ defined in equation (2.3) has all roots outside the unit circle. It is assumed that the roots of $\alpha(z)$ and $\beta(z)$, where

$$\beta(z) = 1 + \beta_1 z + \ldots + \beta_q z^q, \tag{2.6}$$

do not have a common factor. The autocorrelations satisfy

$$\begin{aligned}
\rho_k + \sum_{j=1}^{p} \alpha_j \rho_{|k-j|} &= \frac{1}{\mathrm{var}(y)} \sum_{i=k}^{q} \beta_i E\left[y_{t-k}\varepsilon_{t-i}\right] \qquad 1 \leq k \leq q \\
&= 0 \qquad\qquad k > q.
\end{aligned} \tag{2.7}$$

To evaluate the right hand side of this equation, we need to use the Wold decomposition theorem, which states that if $\{y_t\}$ is a discrete stationary process with finite variance, then there are two mutually uncorrelated

processes $\{\psi_t\}$, $\{\eta_t\}$ such that $y_t = \psi_t + \eta_t$, where

1. ψ_t is deterministic
2. η_t is a (purely non-deterministic) moving average
$$\eta_t = \sum_{j=o}^{\infty} b_j \varepsilon_{t-j} \qquad b_o = 1$$
where $\sum_{j=o}^{\infty} b_j^2 < \infty$ and the ε_t are a sequence of uncorrelated random variables as described above.

The processes ψ_t, η_t are uniquely specified, and either one may be absent.

Thus every stationary process can be written as an infinite moving average, possibly plus a deterministic term. In particular, the purely non-deterministic ARMA(p,q) model has the representation

$$y_t = \sum_{j=o}^{\infty} b_j \varepsilon_{t-j}$$

where the b_j are uniquely specified and $b_o = 1$. We define

$$B(z) = \sum_{i=o}^{\infty} b_i z^i \,. \tag{2.8}$$

Writing y_{t-k} in this form, $E(y_{t-k}, \varepsilon_{t-i})$ becomes $E\left[\varepsilon_{t-i} \sum_{j=o}^{\infty} b_j \varepsilon_{t-k-j}\right]$, so that equation (2.7) can be written

$$\rho_k + \sum_{j=1}^{p} \alpha_j \rho_{|k-j|} = \frac{\mathrm{var}(\varepsilon)}{\mathrm{var}(y)} \sum_{i=k}^{q} \beta_i b_{i-k} \qquad 1 \leq k \leq q \tag{2.9}$$

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

If all the roots of $\beta(z) = 1 + \beta_1 z + \ldots + \beta_q z^q$ are greater than one in modulus then the process also has an infinite autoregressive representation

$$\sum_{i=o}^{\infty} a_i y_{t-i} = \varepsilon_t$$

where $a_o = 1$. When this representation exists, the model is called invertible. We write $A(z) = \sum_{i=o}^{\infty} a_i z^i$. It is well known that when a stationary process is invertible, this generating function and (2.8) are related by $A(z)B(z) = 1$.

It is clear that for the finite MA(q) model, the γ_k and hence the ρ_k are easily found from the β_j. To obtain γ_k for the AR and ARMA models is more difficult. It is possible to invert the Yule-Walker equations, but this is again difficult for ARMA (p,q) models if q is moderately large. An alternative generating function approach is given in an algorithm due to Quenouille (1947a) First, we need to define the autocovariance generating function

$$\Gamma(z) = \gamma_o + \sum_{k=1}^{\infty} \gamma_k (z^k + z^{-k}) \qquad (2.10)$$

which satisfies

$$\Gamma(z) = V(\varepsilon)B(z)B(z^{-1}) = V(\varepsilon)/A(z)A(z^{-1}). \qquad (2.11)$$

Then considering first the AR(p) model and applying Quenouille's algorithm, the expression $\Gamma(z)/V(\varepsilon)$ is uniquely determined by

$$\frac{1}{\alpha(z)\,\alpha(z^{-1})} = K_o + \frac{(K_1 z + K_2 z^2 + \ldots + K_p z^p)}{\alpha(z)} + \frac{(K_1 z^{-1} + \ldots + K_p z^{-p})}{\alpha(z^{-1})} \qquad (2.12)$$

where $K_o, K_1, \ldots K_p$ are found by equating powers of $z^o, z^1, \ldots z^p$. Comparing equations (2.10) and (2.12) we find that

$$\gamma_o = K_o V(\varepsilon)$$

$$\alpha(z) \sum_{k=1}^{\infty} \gamma_k z^k = (K_1 z + K_2 z^2 + \ldots + K_p z^p) V(\varepsilon).$$

For the ARMA(p,q) model, the generating function (2.8) is $B(z) = \beta(z)/\alpha(z)$ and the generalisation of Quenouille's algorithm states that $\Gamma(z)/V(\varepsilon)$ is given by

$$\frac{\beta(z)\beta(z^{-1})}{\alpha(z)\alpha(z^{-1})} = K_o + (K_1 z + \ldots K_L z^L)/\alpha(z) + (K_1 z^{-1} + \ldots K_L z^{-L})/\alpha(z^{-1}) \tag{2.13}$$

where $L = \max(p,q)$.

Example 2.1

The ARMA (1,2) model is given by

$$y_t + \alpha y_{t-1} = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} \quad .$$

Using equation (2.13)

$$\frac{(1+\beta_1 z+\beta_2 z^2)(1+\beta_1 z^{-1}+\beta_2 z^{-2})}{(1+\alpha z)(1+\alpha z^{-1})} = K_o + \frac{(K_1 z + K_2 z^2)}{1+\alpha z} + \frac{(K_1 z^{-1} + K_2 z^{-2})}{1+\alpha z^{-1}} .$$

Multiplying this equation by $(1+\alpha z)(1+\alpha z^{-1})$ and then equating powers of z we obtain

$$K_2 = \beta_2$$

$$K_o(1 + \alpha^2) + 2K_1\alpha = 1 + \beta_1^2 + \beta_2^2$$

$$K_o\alpha + K_1 = \beta_1 + \beta_1\beta_2 - \beta_2\alpha$$

which has the solution

$$K_o(1 - \alpha^2) = 1 + \beta_1^2 + \beta_2^2 - 2\alpha(\beta_1 + \beta_1\beta_2 - \beta_2\alpha)$$

$$K_1(1 - \alpha^2) = (1 + \alpha^2)\beta_1(1 + \beta_2) - \alpha\beta_2 - \alpha(1 + \beta_1^2 + \beta_2^2) - \alpha^3\beta_2$$

from which

$$\gamma_o = K_o V(\varepsilon)$$

$$\gamma_1 = K_1 V(\varepsilon)$$

$$\gamma_k = (K_2 - \alpha K_1)(-\alpha)^{k-2} V(\varepsilon) \qquad k \geq 2 .$$

We now turn to the converse problem that of finding the model parameters from the autocorrelations, which is perhaps more frequently encountered in practice. For the autoregressive model of order p, the solution is straightforward. From the Yule-Walker equations

$$\begin{bmatrix} \rho_1 \\ \vdots \\ \rho_p \end{bmatrix} + \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{p-1} \\ \vdots & & & \vdots \\ \rho_{p-1} & \cdots & & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (2.14)$$

so that

$$\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} = - \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & & \rho_{p-2} \\ \vdots & & \ddots & \vdots \\ \rho_{p-1} & \cdots & & 1 \end{bmatrix}^{-1} \begin{bmatrix} \rho_1 \\ \vdots \\ \rho_p \end{bmatrix} . \qquad (2.15)$$

Equation (2.15) shows that $\alpha_1,\ldots,\alpha_p$ are readily obtained from $\rho_1,\ldots,\rho_p$. The solutions for the moving average and ARMA models are more complicated, and require the Cramér-Wold factorization. Consider the moving average model of order q. Then

$$z^q\Gamma(z) = K\beta(z)z^q\beta(z^{-1}) \qquad (2.16)$$

where K is independent of z. If z_i is a root of $z^q\Gamma(z)$, then by the symmetry of the representation, z_i^{-1} is also a root. Thus $z^q\Gamma(z)$ can be expressed

$$z^q\Gamma(z) = K\beta_q \prod_{i=1}^{q} (z-z_i)(z-z_i^{-1})$$

which means there may be as many as 2^q possible solutions. In practice, it is usually required that the model is invertible, so that all the roots of $\beta(z)$ are greater than one in modulus. This defines the β's uniquely. Methods for factorising (2.16) are discussed in Godolphin (1976a). A matrix approach is described by Tunnicliffe-Wilson (1969).

The corresponding problem for ARMA models is treated similarly, by first transforming the problem to that for a moving average model and then applying the above approach. Details for both MA and ARMA models can be found in Godolphin (1976a).

The spectral density function of a time series $\{y_t\}$ with autocovariance function $\{\gamma_k\}$ is defined by

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_k e^{-i\lambda k} \qquad -\pi < \lambda \leq \pi$$

$f(\lambda)$ is an even function of λ, and is continuous, positive, real valued and periodic with period 2π. It is possible to consider the prediction problem in terms of the spectral density, and to formulate some of the problems of this thesis accordingly. However, this approach is only valid for stationary models, so it will not apply to the non-stationary models we shall consider. Hence we shall make little use of the spectral density in this work.

2.3 Non-stationary Time Series ARIMA models.

The stationary processes described in Section 2.2 have proved to be unacceptable for modelling many time series that occur in practice. Consequently, many attempts have been made to modify the concept of a stationary model to be more in keeping with observed time series and yet keep some of the properties of stationarity. Possibly the simplest of such modifications is the deterministic trend with stochastic stationary disturbance model

$$y_t = m_t + x_t . \qquad (2.17)$$

This appears to be essentially an extension of the representation used in the Wold decomposition theorem, for m_t is a deterministic expression, perhaps a polynomial or a cycle or an asymptotic expression, or any combination of these three, while the stochastic element X_t is a stationary time series. Often, X_t is assumed to be just the purely random process ε_t. The most successful workers in this field appear to be R.G. Brown and his co-authors. The works of Brown (1959, 1962), Holt (1957),Winters (1960), Brown and Meyer (1961) and D'Esopo (1961) introduced the topics of exponential smoothing and exponentially weighted regression to fit the deterministic component m_t.

Whittle (1963, Chapter 8) suggested that a more realistic approach to non-stationary time series is given by a model which becomes stationary on the application of a suitable transformation. In particular, Whittle proposed that differencing was a suitable transformation. This idea, which was first considered by Yaglom (1955) is at the heart of the work of Box and Jenkins (1970). Their approach is now widely used for modelling time series data. The model, known as the autoregressive integrated moving-average model of order (p,d,q), usually abbreviated to ARIMA (p,d,q) is given by

$$\nabla^d y_t+\alpha_1\nabla^d y_{t-1}+\ldots+\alpha_p\nabla^d y_{t-p} = \varepsilon_t+\beta_1\varepsilon_{t-1}+\ldots+\beta_q\varepsilon_{t-q} \qquad (2.18)$$

where $\nabla y_t = y_t - y_{t-1}$ and it is assumed that the polynomials $\alpha(z)$, $\beta(z)$ defined by equations (2.3) and (2.6) have no common zeros, and have all zeros strictly outside the unit circle. This is equivalent to stating

that differencing the data d times leads to an ARMA(p,q) model, for which the theory of the previous section applies. The model (2.18) has been described in three different ways by Box and Jenkins (1970).

The most widely used representation is the difference equation form given by equations (2.18). The inverted form

$$Y_t = \sum_{j=1}^{\infty} \Pi_j Y_{t-j} + \varepsilon_t \qquad (2.19)$$

always exists because we have assumed the process in $\nabla^d Y_t$ to be invertible. This representation can sometimes be useful for describing predictor weights in forecasting situations. Box and Jenkins also define a random shock representation

$$Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$$

but this form seems doubtful because the ψ_j do not necessarily decrease as $j \to \infty$. This 'Wold representation' was criticized by Godolphin (1975).

2.4 Estimation of Parameters of an ARMA process

The Box-Jenkins approach requires that a model is specified in terms of unknown parameters, which are then estimated from the data. Substantial literature exists on the estimation of parameters for stationary models, but there appears to be no similarly detailed theory of estimation for non-stationary models. However, Box-Jenkins models do not require such theories, for they can first be transformed to stationary models by differencing. In this section, we assume that the time series $\{y_t\}$ is a stationary ARMA(p,q) process, with p and q known,

and the ε_t are independent and identically distributed normal random variables. We consider the problem of estimating $\alpha_1,\ldots\alpha_p$, $\beta_1,\ldots\beta_q, V(\varepsilon)$.

The special case $q = 0$ (purely autoregressive process) was considered by Mann and Wald (1943). For large samples the maximum likelihood solutions for $\hat{\alpha}_1,\ldots\hat{\alpha}_p$ are found by the intuitively reasonable approach of solving the Yule-Walker equations (2.4), with α_k replaced by $\hat{\alpha}_k$ and ρ_k replaced by the sample serial correlation r_k defined by

$$r_k = C_k/C_o \qquad (2.20)$$

where C_k is the sample serial covariance of lag k

$$C_k = \frac{1}{N-k} \sum_{t=1}^{N-k} y_t y_{t+k}. \qquad (2.21)$$

With these substitutions, the solution is given by equations (2.15). These were adapted by Durbin (1960) to yield a stepwise method of estimating $\hat{\alpha}_1,\ldots\hat{\alpha}_k$ from $k-1$ estimates derived from a previous step $(k = 2,\ldots p)$. This method is particularly useful if p is unknown. These estimates were proved consistent by Mann and Wald (1943) and efficient in large samples by Whittle (1953).

The other special case, that of the MA(q) model was considered by Whittle (1951, 1953) from the maximum likelihood principle. His solution, which is consistent and efficient, but not in closed form, can be found by an iterative process. A direct representation of the iterative solution in terms of the sample serial correlations has been found by Godolphin (1977, 1978) and a computer implementation published by Angell and Godolphin (1978).

The generalisation of the maximum likelihood approach

to the ARMA(p,q) process was also considered by Whittle (1953, 1954). He derived the variance-covariance matrix $\underline{\Sigma}$ of the maximum likelihood parameter estimators $\hat{\alpha}_1,\ldots\hat{\alpha}_p$, $\hat{\beta}_1,\ldots\hat{\beta}_q$ and concluded that these estimators are efficient, in that the generalised variance $\det \underline{\Sigma}$ is smaller than the corresponding generalised variance of any other set of estimators. An alternative method of computing the elements of $\underline{\Sigma}$ has also been given by Box and Jenkins (1970, Chapter 7).

Expressions for the maximum likelihood estimators for the ARMA(p,q) process have been given by Pham-Dinh (1979) and Godolphin (1980b). The log likelihood for the observations $\underline{Y} = (Y_1,\ldots Y_N)^T$ is given by

$$\log L = -\tfrac{1}{2}\{N\log 2\pi\sigma^2 + \log\det\underline{\Gamma}_N + \sigma^{-2}\underline{Y}^T\underline{\Gamma}_N^{-1}\underline{Y}\} \qquad (2.22)$$

where $\underline{\Gamma}_N$ is the covariance matrix of $\underline{Y}$. Differentiating (2.22) with respect to σ^2 and θ, where θ is any element of $(\alpha_1,\ldots\alpha_p,\ \beta_1,\ldots\beta_q)$, we obtain the following approximaton to the likelihood equations

$$\frac{\partial}{\partial\theta}\ (\underline{Y}^T\underline{\Gamma}_N^{-1}\underline{Y}) = 0 \qquad (2.23)$$

and

$$\hat{\sigma}^2 = \underline{Y}^T\underline{\Gamma}_N^{-1}\underline{Y}/N \qquad (2.24)$$

where the contribution of $\det\underline{\Gamma}_N$ has been ignored. To obtain the solution for $\hat{\underline{\alpha}} = (\hat{\alpha}_1,\ldots\hat{\alpha}_p)^T$ and $\hat{\underline{\beta}} = (\hat{\beta}_1,\ldots\hat{\beta}_q)^T$ to (2.23), it is possible to adopt a further approximation, originally proposed by Whittle. We replace $\underline{\Gamma}_N^{-1}$ by $\underline{\Pi}_N$, where $\underline{\Pi}_N$ is the covariance matrix for N consecutive values of the stationary ARMA(q,p) process $\{z_t\}$

$$z_t + \beta_1 z_{t-1} + \ldots + \beta_q z_{t-q} = \eta_t + \alpha_1\eta_{t-1} + \ldots + \alpha_p\eta_{t-p}$$

where $n_t \sim N(0,1)$ (see Shaman, (1976)). Using this approximation, the likelihood equations (2.22) become

$$\frac{\partial}{\partial \theta} \quad \underline{Y}^T \underline{\pi}_N \underline{Y} = 0$$

which simplify to

$$\frac{\partial}{\partial \theta} \{\pi_o + 2 \sum_{j=1}^{m} \pi_j r_j\} = 0$$

where r_j is defined by equation (2.20) and m is sufficiently large.

Solutions for the likelihood equations can be expressed as iterative equations for $\hat{\underline{\alpha}}$ and $\hat{\underline{\beta}}$ together with the non-iterative solution

$$\hat{\sigma}^2 = N^{-1} \sum_{t=1}^{N} y_t^2 (\pi_o + 2 \sum_{j=1}^{m} \pi_j r_j).$$

Pham-Dinh has also given an approach which enables the iterative procedure to converge quadratically.

Several other attempts have been made to estimate the parameters of MA or ARMA processes. Durbin (1959) approximated the stationary MA process by a large but finite autoregression, and then used the approach of Mann and Wald. This procedure is prone to bias, but is shown by Bhansali (1980) to be asymptotically efficient in estimation of moving average parameters, relative to the maximum likelihood procedure. It also appears to be the only non-iterative procedure. The generalisation of this approach to the mixed ARMA model is given by Durbin (1960).

Another interesting approach is that of Walker (1961, 1962) who estimates not the α's and β's but the α's and ρ's taking the sample serial correlation r_k as the initial estimate of ρ_k. The β's are then found from the Cramér-Wold factorization. This approach requires

the theory of the distribution of the r_k, due to Bartlett (1946) and Lomnicki and Zaremba (1957).

Other methods are due to Anderson (1975a), Hannan (1969) and Box and Jenkins (1970). These last two are computational procedures for obtaining the maximum likelihood estimator directly.

2.5 Specification

If the observed time series is non-stationary, it may be decided to fit an ARIMA model. The first step is to determine d, the number of times the data should be differenced, and this must be done before estimating any other parameters. The method suggested by Box and Jenkins (1970) to do this is to compute the sample serial autocorrelations r_k defined in equation (2.20). When the data are stationary and N is large, the sample serial correlation function $\{r_k, k=1,2,...\}$ follows closely the behaviour of the theoretical autocorrelation function $\{\rho_k, k=1,2,...\}$, indeed asymptotically $E(r_k) = \rho_k$ (Lomnicki and Zaremba 1957). For this reason it is assumed that the degree of differencing, d, necessary to achieve stationarity has been reached when the estimated autocorrelation function of $w_t = \nabla^d y_t$ dies away fairly quickly. See Box and Jenkins (1970, P.174-175) for a fuller discussion. A plot of the correlogram (r_k against k) is usually helpful.

Having achieved a stationary ARMA (p,q) model by this process, it seems reasonable that p and q should be determined before estimating the α's and β's However, many practicioners tend to overfit a stationary model,

in the hope that estimates insignificantly different from zero will effectively determine the order of the process. For example, in applying Durbin's method for the estimation of parameters of an autoregressive process, p is taken to be the number of parameters significantly different from zero. In general, $V(\varepsilon)$ can be estimated at each iteration, and only minor reductions in $\hat{V}(\varepsilon)$ imply that estimating further parameters will not improve the fit.

Akaike (1974b) attempts to formalize this idea in the autoregressive case, assuming the ε_t are normally distributed. He defines an information criterion

$$AIC(p) = N \log \hat{V}(\varepsilon) + 2p$$

where

$$\hat{V}(\varepsilon) = \sum_{i=o}^{p} \hat{a}_i C_i$$

and the $\hat{a}_i$ are the solution of the Yule-Walker equations (2.4) with ρ_k replaced by the sample serial correlation r_k, and C_k is the sample serial covariance defined in (2.21). The order of the model, p, is taken as the value of p which minimizes $AIC(p)$ for $p = 0,1,\ldots L$, where L is a preassigned upper limit. A Bayesian approach is also described by Akaike (1979). A critical examination of Akaike's method is given by Bhansali and Downham (1977) and Shibata (1976).

Several authors have proposed methods for testing the goodness of fit of a stationary ARMA model. Quenouille (1947b) provided a test of fit of an autoregressive process, using partial autocorrelations. This was extended by Bartlett and Diananda (1950). Walker (1952) compares the

power of these two tests. Wold (1949) and Durbin(1959) give goodness of fit tests for a moving average process which yields a χ^2 statistic under the null hypothesis. Durbin's test was generalised to ARMA models by Durbin (1960).

One test which is reported to have good power properties is due to Whittle (1951, 1952) and is based upon the likelihood ratio principle. In practice, however, a simple test which is often used is the Box-Pierce test (1970), or its modification by Ljung and Box (1978). The Box-Pierce test requires the computation of

$$N \sum_{j=1}^{T} r_j^2$$

where r_j is the sample serial correlation defined by equation (2.20) and T is a sufficiently large integer less than N. This statistic has a limiting χ^2 distribution on $T-p-q$ degrees of freedom. However, this simple test has the reputation of being unable to distinguish between several models which could be fitted to the data. The modification of Ljung and Box (1978) gives a test statistic

$$N(N+2) \sum_{k=1}^{T} (N-k)^{-1} r_k^2$$

which is a closer approximation to χ^2 on $T-p-q$ degrees of freedom. Davies and Newbold (1979) have compared these two tests with regard to forecasting accuracy.

Another approach which contains the Box-Pierce test as a special case has been proposed by Godolphin (1980a). This method requires rather more computation than the Box-Pierce test, but has greater power properties. It is based on Walker's idea that we should test the ρ's rather than the β's, using the Godolphin (1978) estimation procedure described in section 2.4. The set of sample

serial correlations $(r_1,\ldots,r_T)$ are transformed to a set $\underline{w} = (w_1 \ldots w_{T-k})$, which is partitioned into $(\underline{w}_*, w_{m+1}, \ldots w_{T-k})$, with transformed covariance matrix

$$\begin{bmatrix} \underline{\Omega}_{11} & \underline{\Omega}_{12} \\ \underline{\Omega}_{12}^T & \underline{\Omega}_{22} \end{bmatrix} .$$

$\underline{w}_* = (w_1,\ldots w_m)^T$ is then estimated by Walker's iterative procedure. The test statistic

$$NQ_{T-k-m} = N\underline{w}_*^T(\underline{\Omega}_{11}-\underline{\Omega}_{12}\underline{\Omega}_{22}^{-1}\underline{\Omega}_{12}^T)^{-1}\underline{w}_*$$

has a central χ^2 distribution on m degrees of freedom under the null hypothesis that the parameters of the ARMA model have been correctly specified.

2.6 Fitting models to data

Several papers have been written which attempt to apply the theory of Sections 2.2 - 2.5 to real time series data. For example, Chatfield and Prothero (1973) consider one data series in some detail, while Newbold and Granger (1974) and Prothero and Wallis (1976) both compare the Box-Jenkins models to other models over several data sets. Several papers describing the application of the Box-Jenkins methods to specialised problems have now appeared in the applied journals, suggesting that the use of these methods is widespread. In general, p and q are found to be quite small, generally not more than 3.

An attempt at fitting a deterministic term plus a stationary model to the so called lynx data has been made by Campbell and Walker (1976). These data have also been examined by Tong (1976) who fitted an AR model based on Akaike's criterion.

CHAPTER 3

PREDICTION OF TIME SERIES MODELS

3.1 Introduction

In this chapter, we are concerned with the prediction of linear time series models defined in Chapter 2; that is, we wish to predict y_{t+k} $k > 0$ from the data available at time t. This is clearly useful in many contexts to enable decisions to be made, for example in sales forecasting.

Methods of prediction have occupied many workers, but some of these methods appear not to interest practitioners, for example, the work of Yaglom (1962, Chapter 6) and the general form of what is now known as the Wiener-Kolmogorov approach.

The approaches discussed in this chapter are due to several authors. The material of Section 3.2 is discussed in Whittle (1963) and Box and Jenkins (1970), where references to relevant workers can be found. The major source of the material of Section 3.3 is Godolphin (1975).

These approaches assume that the model is known. In practice, it is first necessary to specify the order of the ARMA or ARIMA process, and estimate the parameters. In the sequel, we shall assume this has been done.

3.2 Prediction of Stationary Time Series

We assume that the time series $\{y_t\}$ is stationary and non-deterministic. Then by the Wold decomposition theorem, y_t has an infinite moving average representation

$$y_t = \sum_{j=o}^{\infty} b_j \varepsilon_{t-j}$$

with $b_o = 1$. It is required to predict y_{t+k} at time t from a linear combination of past and present values of $\{y_t\}$. Thus the predictor $y_t(k)$ of y_{t+k} is given by

$$y_t(k) = \sum_{j=o}^{\infty} q_j(k) y_{t-j} \cdot \qquad (3.1)$$

We wish to choose the weights $q_j(k)$ such that the prediction error variance $E\left[(y_t(k) - y_{t+k})^2\right]$ is minimised. Minimising this quantity gives

$$y_t(k) = \sum_{j=o}^{\infty} b_{j+k} \varepsilon_{t-j} \cdot \qquad (3.2)$$

This result seems reasonable, since the predictor is the expectation of y_{t+k} conditional on $y^t = (y_t, y_{t-1}, \ldots)$, i.e. $y_t(k)$ is obtained from the expression for y_{t+k} by setting future values of $\{\varepsilon_t\}$ to their expected value, zero. The notation y^t for $(y_t, y_{t-1}, \ldots)$ has been adopted from Harrison and Stevens (1976).

Also, the prediction error $y_{t+k} - y_t(k)$ is a finite moving average

$$y_{t+k} - y_t(k) = \varepsilon_{t+k} + \beta_1 \varepsilon_{t+k-1} + \ldots + \beta_{k-1} \varepsilon_{t+1},$$

with prediction error variance

$$\text{var}(\varepsilon) \sum_{i=o}^{k-1} b_i^2,$$

which increases monotonically with k.

However, the formulation (3.2) for the predictor is not very useful because the ε_t are not observed. If we write

$$y_t(k) = \sum_{j=o}^{\infty} \psi_j(k) \varepsilon_{t-j}$$

and define the generating functions

$$\Psi_k(z) = \sum_{j=o}^{\infty} \psi_j(k) z^j \quad \text{and} \quad Q_k(z) = \sum_{j=o}^{\infty} q_j(k) z^j$$

then the relationship between these functions is given by

$$\Psi_k(z) = B(z) Q_k(z).$$

Hence

$$Q_k(z) = \sum_{j=o}^{\infty} b_{j+k} z^j \Big/ \sum_{j=o}^{\infty} b_j z^j \,. \tag{3.3}$$

The $q_j(k)$ can be determined from (3.3) so that the predictor $y_t(k)$ can be found in terms of the previous y_s from (3.1).

It should be noted that the $q_j(k)$ can only be determined if (3.3) can be expressed as a power series in z. For the three models discussed in Section 2.2, $B(z) = \beta(z)/\alpha(z)$. Hence it is clear from (3.3) that the roots of $\beta(z)$ must be greater than one in modulus, i.e. if a predictor of the form (3.1) is to exist, then the model must be invertible.

Example 3.1

Consider a first order autoregressive model

$$y_t - \rho y_{t-1} = \varepsilon_t \,.$$

Then

$$B(z) = \sum_{j=o}^{\infty} \rho^j z^j \,,$$

hence

$$Q_k(z) = \sum_{j=o}^{\infty} \rho^{j+k} z^j \Big/ \sum_{j=o}^{\infty} \rho^j z^j = \rho^k.$$

Thus

$$y_t(k) = \rho^k y_t \,.$$

In this case, the predictor for any lead time depends only on the last observation, and since $|\rho| < 1$, it converges to zero for large lead times.

Relation between predictor weights and autocorrelations

Another approach is to examine the relation of the predictor to the autocorrelation function $\{\rho_j\}$. Here it is assumed that the available data set is of finite length. Suppose $Y_t, Y_{t-1}, \ldots, Y_{t-N}$ is available, and define

$$\underline{G}_N = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdot & \cdot & \cdot & \rho_N \\ \rho_1 & 1 & \rho_1 & & & & \rho_{N-1} \\ \cdot & & \cdot & & & & \cdot \\ \cdot & & & \cdot & & & \cdot \\ \cdot & & & & \cdot & & \cdot \\ \cdot & & & & & \cdot & \cdot \\ \rho_N & \rho_{N-1} & \rho_{N-2} & \cdot & \cdot & \cdot & 1 \end{bmatrix}$$

$\underline{G}_N$ is positive definite and hence invertible. As before, we wish to define the predictor of Y_{t+k} as a linear combination of past and present observations, i.e.

$$Y_t(k) = \sum_{j=o}^{N} q_j(k) Y_{t-j} .$$

Notice that we keep the same notation as above, though the values taken may differ because of the finite approximation to the infinite series.

Defining

$$\underline{Y}_N = \left[Y_t, Y_{t-1}, \ldots Y_{t-N}\right]^T, \quad \underline{Q}_k = \left[q_o(k), \ldots q_N(k)\right]^T$$

we have

$$Y_t(k) = \underline{Q}_k^T \underline{Y}_N . \tag{3.4}$$

Minimising the prediction error variance $E\left[(Y_{t+k} - Y_t(k))^2\right]$, we find that

$$\underline{G}_N \underline{Q}_k = \underline{\rho}_k \tag{3.5}$$

where

$$\underline{\rho}_k = \left[\rho_k, \rho_{k+1}, \ldots \rho_{k+N}\right]^T .$$

Since $\underline{G}_N$ is positive definite, there is a unique solution for $\underline{Q}_k$, hence $y_t(k)$ is uniquely defined by (3.4). The prediction error variance is given by

$$E\left[(y_{t+k}-y(k))^2\right] = \gamma_o(1-\underline{\rho}_k^T\underline{Q}_k) \ .$$

Inverting $\underline{G}_N$ may be difficult, but sometimes it is possible for a solution to be hypothesised. If this solution fits equation (3.5), then it must be the unique solution.

Prediction by the spectral density function

It is also possible to predict from the spectral density function. This involves factorisation of the spectral density to find $B(z)$, and then proceed in the manner described earlier to give equation (3.3) as before.

3.3. Prediction of ARIMA models

Since an ARIMA (p,d,q) process is not stationary, the predictors cannot be defined from the Wold canonical form as in Section 3.2. However, the predictor of y_{t+k} at time t can still be defined as the expectation of y_{t+k} conditional on y^t as in the previous section. Thus we replace t by $t+k$ in equation (2.18) and take expectations conditional on y^t using

$$
\begin{aligned}
E\left[y_{t+j}|y^t\right] &= y_t(j) && 1 \leq j \leq k \\
E\left[\varepsilon_{t+j}|y^t\right] &= 0 && 1 \leq j \leq k \\
E\left[y_{t-j}|y^t\right] &= y_{t-j} && 0 \leq j \leq p+d-k \\
E\left[\varepsilon_{t-j}|y^t\right] &= y_{t-j}-y_{t-j-1}(1) && 0 \leq j \leq q-k \ .
\end{aligned}
\tag{3.6}
$$

This procedure results in a difference equation in the predictors $y_t(j)$, $y_{t-j}(1)$, and the observations y_{t-j}, which can be solved to find $y_t(k)$ in terms of the y_{t-j}. Again, we use the same notation, although the predictor $y_t(k)$ is now defined by conditional expectation, not minimum mean square error. Unless otherwise stated, predictor will be taken to mean conditional expectation predictor defined by (3.6).

Example 3.2

The ARIMA (0,1,1) model is given by

$$y_t - y_{t-1} = \varepsilon_t + \beta\varepsilon_{t-1} \qquad \text{where } |\beta| < 1 \ .$$

Replacing t by $t+1$ and taking conditional expectations as in (3.6),

$$y_t(1) + \beta y_{t-1}(1) = (\beta+1)y_t .$$

This has solution

$$y_t(1) = (1+\beta) \sum_{j=0}^{\infty} (-\beta)^j y_{t-j} \tag{3.7}$$

and for $k \geq 2$

$$y_t(k) - y_t(k-1) = 0$$

$$\text{i.e. } y_t(k) = y_t(1) \text{ for all } k \geq 1 \ . \tag{3.8}$$

Thus the predictor of this model is the same for all lead times. In this work, a model whose predictor satisfies

(3.8) is called a steady model. The expression (3.7) is known as the exponentially weighted moving average (EWMA), and is widely used in industrial forecasting. It should be noted that the condition $|\beta| < 1$ is a natural one for the representation (3.7) to be valid.

In general, the difference equation is very difficult to solve. Godolphin (1975) produced what he called the direct basic form, which produces forecasts of y_{t+k} without requiring forecasts of $y_{t+1},\ldots,y_{t+k-1}$, and is fairly straightforward to find if p is small. It is convenient to consider first the case $p = 0$.

Given an ARIMA (0,d,q) model, we define the updating series

$$U_t = (1+\beta_1)y_t + \sum_{j=1}^{q-1} \beta_{j+1} y_{t-j} - \sum_{j=1}^{q} \beta_j U_{t-j} \qquad (3.9)$$

and the component series

$$C_t = y_t - U_{t-1} \qquad (3.10)$$

from which

$$U_t = y_t + \sum_{j=o}^{q-1} \beta_{j+1} C_{t-j} \,. \qquad (3.11)$$

Then we have

$$y_t(1) = U_t + \sum_{i=o}^{d-2} \nabla^i C_t \qquad (3.12)$$

where the last term vanishes if d is one. If we define

$$f_j = E\left[\nabla^d y_{t+j} | y^t\right] \qquad j \geq o \qquad (3.13)$$

it can be shown that

$$f_j = \sum_{i=o}^{q-j} \beta_{i+j} \nabla^{d-1} C_{t-i} \qquad 1 \leq j \leq q$$

$$= o \qquad j \geq q+1 \,.$$

From this, the predictor is found to be

$$y_t(k) = \sum_{i=o}^{d-1} \binom{k+i-2}{i} A_{it} \qquad k \geq 2 \qquad (3.14)$$

where

$$A_{it} = \nabla^i \Omega_{d-i} + (-1)^{d-1-i} \sum_{j=2}^{r} \binom{j-2}{d-1-i} f_j \qquad 0 \leq i \leq d-1$$

and

$$r = \min(q, k+d-1).$$

Here

$$\Omega_1 = U_t, \quad \Omega_d = U_t + C_t + \ldots + \nabla^{d-2} C_t, \qquad d \geq 2.$$

Clearly equation (3.14) defines a polynomial in k of degree $d-1$. When $q \leq d$, $r = q$, so that the A_{it} are independent of k, and the $y_t(k)$ describe a polynomial path for all lead times k. When $q > d$, A_{it} does not become independent of k until $k > q-d$. Thus the predictor has $q-d$ discontinuities, then settles to a forward shifted polynomial of degree $d-1$ and shift $q-d$.

Godolphin and Harrison (1975) give an equivalent representation in Lemma 1 of their paper which uses a matrix updating procedure.

Example 3.3. (The Steady Model)

Consider again the ARIMA (0,1,1) model

$$y_t - y_{t-1} = \varepsilon_t + \beta \varepsilon_{t-1} .$$

Applying the above algorithm,

$$U_t = y_t + \beta C_t$$

$$C_t = y_t - U_{t-1}$$

so that

$$U_t + \beta U_{t-1} = (1 + \beta) y_t$$

i.e. U_t is the exponentially weighted moving average.

Also

$$f_j = 0 \qquad j \geq 2 .$$

Thus

$$y_t(1) = U_t$$

and

$$y_t(k) = A_{ot} = \Omega_1 = U_t \qquad \text{for all } k \geq 2$$

thus confirming our previous result.

Example 3.4 (Linear Growth Model)

A natural definition of linear growth is

$$y_t = a + bt.$$

It is easily shown that in this case, $\nabla^2 y_t \equiv 0$. In practice, a true straight line seldom arises, but close approximations, where $\nabla^2 y_t$ is stationary seem to be met frequently. For example, if $\nabla^2 y_t$ is a moving average of order 2, we have an ARIMA (0,2,2) model, which is widely known as the linear growth model:

$$\nabla^2 y_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} .$$

For this model, the above algorithm yields

$$U_t = y_t + \beta_1 C_t + \beta_2 C_{t-1}$$

$$C_t = y_t - U_{t-1}$$

$$f_2 = \beta_2 \nabla C_t$$

$$f_j = 0 \qquad j \geq 3.$$

Applying (3.12) and (3.14),

$$y_t(1) = \Omega_2 = U_t + C_t$$

$$y_t(k) = \binom{k-2}{0} A_{ot} + \binom{k-1}{1} A_{1t}$$

$$= \Omega_2 + (k-1)(\nabla \Omega_1 + f_2)$$

$$= U_t + C_t + (k-1)(\nabla U_t + \beta_2 \nabla C_t)$$

$$= U_t + C_t + (k-1)(1+\beta_1+\beta_2) C_t .$$

Thus the predictor grows linearly, with slope $(1+\beta_1+\beta_2)C_t$.

Example 3.5 (Forward shifted linear growth model)

An example of a forward shifted model is given by the ARIMA (0,2,3) process:

$$\nabla^2 y_t = \varepsilon_t+\beta_1\varepsilon_{t-1}+\beta_2\varepsilon_{t-2}+\beta_3\varepsilon_{t-3} \quad .$$

Applying the above algorithm, we have

$$U_t = y_t+\beta_1 C_t+\beta_2 C_{t-1}+\beta_3 C_{t-2}$$

$$C_t = y_t - U_{t-1}$$

$$f_2 = \beta_2\nabla C_t+\beta_3\nabla C_{t-1}$$

$$f_3 = \beta_3\nabla C_t$$

$$f_j = 0 \qquad\qquad j \geq 4.$$

From this,

$$y_t(1) = U_t + C_t$$

$$y_t(k) = A_{ot} + (k-1)A_{1t}$$

$$= U_t + C_t - f_3 + (k-1)(\nabla U_t+f_2+f_3)$$

$$= U_t + C_t - \beta_3\nabla C_t + (k-1)(1+\beta_1+\beta_2+\beta_3)C_t \qquad k \geq 2 \ .$$

Thus $y_t(k)$ defines a straight line of slope $(1+\beta_1+\beta_2+\beta_3)C_t$ for $k \geq 2$. Clearly, there is a discontinuity of $(-\beta_3\nabla C_t)$ between $y_t(1)$ and $y_t(2)$, illustrating the reason for the term of forward-shifted polynomial model.

Godolphin (1975) generalised this approach to the case $p \neq 0$, and also to seasonal processes. Given an ARIMA (p,d,q) model, the expressions for the updating and component series (3.10) and (3.11) remain unaltered, as

does the definition (3.13) of f_j. The one step ahead predictor is now given by

$$y_t(1) = U_t + \sum_{i=o}^{d-2} \nabla^i C_t - \sum_{j=1}^{p} \alpha_j \nabla^{d-1} C_{t+1-j}$$

where the central term vanishes when $d = 1$. It can be shown that

$$f_1 = \sum_{j=o}^{q-1} \beta_{j+1} \nabla^{d-1} C_{t-j} - \sum_{i=o}^{p-1} \alpha_{i+1} \nabla^{d-1} C_{t-i}$$

$$f_j = \sum_{L=o}^{q-j} \beta_{L+j} \sum_{i=o}^{p} \alpha_i \nabla^{d-1} C_{t-i-L} - \sum_{i=o}^{p-j} \alpha_{i+j} \nabla^d y_{t-i}$$

$$- \sum_{i=1}^{j-1} \alpha_i f_{j-i} \qquad j = 2,\ldots,q \qquad (3.15)$$

and subsequent f_j satisfy

$$f_j + \alpha_1 f_{j-1} + \ldots + \alpha_p f_{j-p} = 0 \qquad j \geq q+1$$

where, if $j > p$, the second term of equation (3.15) vanishes, and the upper limit on the final summation changes from $j-1$ to p.

From this, the k-step ahead predictor is found to be

$$y_t(k) = \sum_{i=o}^{d-1} \binom{k+i-2}{i} \nabla^i \Omega_{d-i} + \sum_{j=o}^{k-2} \binom{d+j-1}{j} f_{k-j} \qquad (k \geq 2)$$

where

$$\Omega_1 = U_t - \sum_{i=1}^{p} \alpha_i \nabla^{d-1} C_{t+1-i}$$

$$\Omega_d = U_t + \sum_{i=o}^{d-2} \nabla^i C_t - \sum_{j=1}^{p} \alpha_j \nabla^{d-1} C_{t+1-j} \qquad d \geq 2 .$$

Example 3.6

The ARIMA (1,1,2) model is given by

$$\nabla y_t + \alpha \nabla y_{t-1} = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} .$$

For this model

$$U_t = y_t + \beta_1 C_t + \beta_2 C_{t-1}$$

$$C_t = y_t - U_{t-1}$$

as for the linear growth model. However

$$f_1 = \beta_1 C_t + \beta_2 C_{t-1} - \alpha C_t$$

$$f_2 = (\beta_2 - \alpha\beta_1 + \alpha^2) C_t$$

$$f_r = (-\alpha)^{r-2} f_2 \qquad (r \geq 3)$$

so that

$$y_t(1) = U_t - \alpha C_t$$

$$y_t(k) = \Omega_1 + \sum_{j=o}^{k-2} f_{k-j}$$

$$= U_t - \alpha C_t + (\beta_2 - \alpha\beta_1 + \alpha^2) C_t \sum_{r=2}^{k} (-\alpha)^{r-2}$$

$$= U_t - \alpha C_t + (\beta_2 - \alpha\beta_1 + \alpha^2)(1 + (-1)^k \alpha^{k-1})(1+\alpha)^{-1} C_t.$$

We see that this process is neither a constant forecast model nor a linear growth model, but is something between the two, for the predictor converges in a geometric fashion towards a constant forecast. Indeed, as $k \to \infty$, $y_t(k)$ tends towards a ceiling value of $U_t - \alpha C_t + \frac{(\beta_2 - \alpha\beta_1 + \alpha^2)}{1+\alpha} C_t$. If α is negative, then the predictor will converge monotonically towards this ceiling forecast, while if α is positive, the predictor will oscillate towards the limit.

3.4 Equivalence theorems for the predictors of non-stationary models

In the previous sections, we have defined predictors of ARMA and ARIMA models. It is quite possible that other univariate models, for example those of Brown (1962) or Holt-Winters (Winters, 1960) or the DLM to be defined in Chapter 4, have the same predictors as the Box-Jenkins models.

Example 3.7

Consider the simplest form of model (2.17)

$$y_t = m_t + \varepsilon_t$$

where ε_t is a completely random process and $m_t = m_{t-1} = m$, i.e. a constant model with observation error. Here $\hat{m}_t$, the estimate of m after t observations, is $\bar{y}_t = \sum_{i=1}^{t} y_i/t$. The k-step ahead predictor $y_t(k)$ is found by taking expectations of y_{t+k} conditional on $y^t = (y_1,\ldots,y_t)$ so that

$$\begin{aligned} y_t(k) &= E\left[m_{t+k}|y^t\right] + E\left[\varepsilon_{t+k}|y^t\right] \\ &= E\left[m|y^t\right] \\ &= \hat{m}_t \qquad \text{for all } k. \end{aligned}$$

Thus we see that this model is a steady model in that (3.8) is satisfied, so that the predictor is the same for all lead times.

In the remainder of this chapter, we consider conditions under which predictors of various types of model are equivalent to those of the Box-Jenkins models. Godolphin and Harrison (1975) proved the following theorem.

Theorem 3.1

Suppose the predictor $\{y_t(k);k=1,2,\ldots\}$ satisfy

$$\sum_{j=0}^{d} (-1)^j \binom{d}{j} y_t(k-j) = 0. \qquad k \geq d+1. \tag{3.16}$$

Then $y_t(k)$ is identical to the k-step ahead predictor of an ARIMA (0,d,q) process, where $q \leq d$, for each $k = 1,2,\ldots$ if and only if there is a real finite sequence $\alpha_0 = 1,\alpha_1,\ldots,\alpha_d$ such that

(i) $y_t(k) = y_{t-1}(k+1) + \alpha_k e_t \qquad 1 \leq k \leq d$ (3.17)

where $e_t = y_t - y_{t-1}(1)$ is the one-step ahead prediction error

(ii) the zeros of $\beta(z) = 1 + \beta_1 z + \ldots + \beta_d z^d$ lie strictly outside the unit circle, where

$$\beta_j = \sum_{i=o}^{j} (-1)^i \binom{d}{i} \alpha_{j-i} \qquad 1 \leq j \leq d . \quad (3.18)$$

Applying this theorem to the above example, we see that

$$y_t(k) - y_t(k-1) = 0 \qquad \text{for } k \geq 2, \text{ hence } d = 1$$

$$y_{t-1}(k+1) = \hat{m}_{t-1} = \sum_{i=1}^{t-1} y_i/t-1$$

$$\hat{m}_t = \sum_{i=1}^{t} y_i/t = \hat{m}_{t-1} + (y_t - \hat{m}_{t-1})/t$$

which is equation (3.17) with $\alpha_k = 1/t$ for all k. But it is implicitly assumed in the theorem that the α_i do not vary with time. As $t \to \infty$, $\alpha_k \to 0$, but this means that $\beta_1 \to -1$, so that in the limit, $\beta(z)$ has a root on the unit circle. Thus the predictor of this model is not at any time equivalent to that of the ARIMA (0,1,1) model.

The result can be extended as follows:

Theorem 3.2

Let $r \geq 0$, $s > 0$.

Suppose the forecast function $\{y_t(k); k=1,2,\ldots\}$ satisfies

$$\sum_{j=o}^{s} \phi_j y_t(k-j) = 0 \qquad k \geq r + s + 1 \quad (3.19)$$

where the ϕ_i are real scalars, with $\phi_o = 1$ and the roots of $\Phi(z) = \sum_{i=o}^{s} \phi_i z^i$ are greater than one in modulus. It is assumed that s is the smallest integer for which

such a representation exists, and that

$$\sum_{j=o}^{s} \phi_j y_t(k-j) \neq 0 \qquad 1 \leq k \leq r+s \ .$$

Then $y_t(k)$ is identical to the k-step ahead predictor of an ARMA (s,r+s) process for each $k = 1,2,\ldots$ if and only if there is a real finite sequence $\alpha_o = 1, \alpha_1, \ldots \alpha_{r+s}$ such that

(i) $y_t(k) = y_{t-1}(k+1) + \alpha_k e_t \qquad 1 \leq k \leq r+s \qquad (3.20)$

where e_t is the one-step ahead prediction error, defined as above

(ii) the zeros of $\beta(z) = 1 + \beta_1 z + \ldots + \beta_{r+s} z^{r+s}$ lie strictly outside the unit circle, where

$$\beta_j = \sum_{i=o}^{j} \phi_i \alpha_{j-i} \qquad 1 \leq j \leq s \ . \qquad (3.21)$$

$$\beta_{j+s} = \sum_{i=o}^{s} \phi_i \alpha_{s+j-i} \qquad 1 \leq j \leq r \ . \qquad (3.22)$$

Proof

From equation (3.20)

$$y_t(r+s) = y_{t-1}(r+s+1) + \alpha_{r+s} e_t \ .$$

Substituting for $y_{t-1}(r+s+1)$ from equation (3.19) we obtain,

$$y_t(r+s) = -\left[\phi_1 y_{t-1}(r+s) + \ldots + \phi_s y_{t-1}(r+1)\right] + \alpha_{r+s} e_t \ . \qquad (3.23)$$

Using equation (3.20) repeatedly, we obtain

$$\begin{aligned} y_t(1) &= y_{t-1}(2) + \alpha_1 e_t \\ &= y_{t-2}(3) + \alpha_2 e_{t-1} + \alpha_1 e_t \\ &= y_{t-m}(m+1) + \sum_{i=1}^{m} \alpha_i e_{t+1-i} \qquad m \geq 1 \ . \end{aligned} \qquad (3.24)$$

In particular,

$$y_t(1) = y_{t+1-s-r}(r+s) + \sum_{i=1}^{r+s-1} \alpha_i e_{t+1-i}$$

and using equation (3.23), this becomes

$$y_t(1) = -\left[\phi_1 y_{t-s-r}(r+s) + \ldots + \phi_s y_{t-s-r}(r+1)\right] + \sum_{i=1}^{r+s} \alpha_i e_{t+1-i} . \qquad (3.25)$$

It is required to find expressions for $y_{t-s-r}(k)$ in terms of $y_{t-i}(1)$ and e_{t+1-i}. To this end, we use equation (3.24) to find

$$\begin{aligned} y_{t-s-r}(r+s) &= y_{t-1}(1) - \sum_{i=1}^{r+s-1} \alpha_i e_{t-i} \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots \\ y_{t-s-r}(r+1) &= y_{t-s}(1) - \sum_{i=1}^{r} \alpha_i e_{t-s+1-i} . \end{aligned} \qquad (3.26)$$

Substituting equations (3.26) into equation (3.25)

$$\begin{aligned} y_t(1) + \sum_{i=1}^{s} \phi_i y_{t-i}(1) &= \sum_{i=1}^{s} \phi_i \sum_{j=1}^{r+s-i} \alpha_j e_{t+1-j-i} + \sum_{i=1}^{r+s} \alpha_i e_{t+1-i} \\ &= \sum_{k=1}^{s} e_{t+1-k} \sum_{j=o}^{k-1} \phi_j \alpha_{k-j} \\ &\quad + \sum_{k=s+1}^{r+s} e_{t+1-k} \sum_{j=o}^{s} \phi_j \alpha_{k-j} . \end{aligned} \qquad (3.27)$$

Since $y_t(1)$ is the one-step ahead forecast $y_{t+1} - e_{t+1}$, (3.27) can be written

$$\begin{aligned} y_{t+1} + \sum_{i=1}^{s} \phi_i y_{t+1-i} &= e_{t+1} + \sum_{i=1}^{s} \phi_i e_{t+1-i} \\ &+ \sum_{k=1}^{s} e_{t+1-k} \sum_{j=o}^{k-1} \phi_j \alpha_{k-j} + \sum_{k=s+1}^{r+s} e_{t+1-k} \sum_{j=o}^{s} \phi_j \alpha_{k-j} \\ &= \sum_{i=o}^{r+s} \beta_i e_{t+1-i} \end{aligned} \qquad (3.28)$$

where

$$\beta_o = 1$$

$$\beta_i = \sum_{j=o}^{i} \phi_j \alpha_{i-j} \qquad 1 \leq i \leq s$$

$$\beta_{s+i} = \sum_{j=o}^{s} \phi_j \alpha_{s+i-j} \qquad 1 \leq i \leq r.$$

Thus interpreting the e_t as purely random variables, equation (3.28) is the conventional form of the ARMA (s,r+s) model, with the β_s given by (3.21) and (3.22). It is, however, essentially a formulation of the one step ahead predictor. We should ensure that the equations for the predictors of all lead times result in the same ARMA (s,r+s) model. In fact, this only needs to be checked for $1 \leq k \leq r+s$, then the results for lead times greater than $r+s$ follow from equation (3.19).

Let $2 \leq m \leq s$, and replce t by $t+m-1$ in equation (3.27) to give

$$Y_{t+m-1}(1) + \sum_{i=1}^{s} \phi_i Y_{t+m-1-i}(1) = \sum_{k=1}^{s} e_{t+m-k} \sum_{j=o}^{k-1} \phi_j \alpha_{k-j}$$

$$+ \sum_{k=s+1}^{r+s} e_{t+m-k} \sum_{j=o}^{s} \phi_j \alpha_{k-j}. \quad (3.29)$$

Since $Y_{t+m-j-1}(1) = Y_t(m-j) + \sum_{i=1}^{m-j-1} \alpha_i e_{t+m-i-j}$ for $j \leq m-2$ from equation (3.24), and invoking (3.22), equation (3.29) can be written

$$Y_t(m)+\phi_1 Y_t(m-1)+\ldots+\phi_{m-2}Y_t(2)+\phi_{m-1}Y_t(1)+\phi_m Y_{t-1}(1)+$$

$$\ldots+\phi_s Y_{t+m-1-s}(1)$$

$$= -\sum_{i=1}^{m-1} \alpha_i e_{t+m-i} - \sum_{j=1}^{m-2} \phi_j \sum_{i=1}^{m-j-1} \alpha_i e_{t+m-i-j}$$

$$+ \sum_{k=1}^{s} e_{t+m-k} \sum_{j=o}^{k-1} \phi_j \alpha_{k-j} + \sum_{k=s+1}^{r+s} e_{t+m-k} \beta_k .$$

Noting that $y_{t-j}(1) = y_{t-j+1} - e_{t-j+1}$ for $j \geq 0$, we have

$$\sum_{i=o}^{m-1} \phi_i y_t(m-i) + \sum_{i=m}^{s} \phi_i y_{t+m-i} = \sum_{i=m}^{s} \phi_i e_{t+m-i} - \sum_{i=1}^{m-1} \alpha_i e_{t+m-i}$$

$$- \sum_{k=2}^{m-1} e_{t+m-k} \sum_{j=1}^{k-1} \phi_j \alpha_{k-j} + \sum_{k=1}^{s} e_{t+m-k} \sum_{j=o}^{k-1} \phi_j \alpha_{k-j} + \sum_{k=s+1}^{r+s} e_{t+m-k} \beta_k$$

and using equation (3.21) we obtain

$$\sum_{i=o}^{m-1} \phi_i y_t(m-i) + \sum_{i=m}^{s} \phi_i y_{t+m-i} = \sum_{k=m}^{r+s} e_{t+m-k} \beta_k \qquad (3.30)$$

which is the difference equation formula specified by Box and Jenkins (1970, p.129).

Similarly, when $2+s \leq m \leq r+s$, the corresponding equation to (3.29) is

$$\sum_{i=o}^{s} \phi_i y_t(m-i) = - \sum_{i=o}^{s} \phi_i \sum_{j=1}^{m-i-1} \alpha_j e_{t+m-i-j}$$

$$+ \sum_{k=1}^{s} e_{t+m-k} \sum_{j=o}^{k-1} \phi_j \alpha_{k-j} + \sum_{k=s+1}^{r+s} e_{t+m-k} \sum_{j=o}^{s} \phi_j \alpha_{k-j}$$

$$= \sum_{k=m}^{r+s} e_{t+m-k} \beta_k \qquad (3.31)$$

and when $m = s+1$, we have

$$\sum_{i=o}^{s} \phi_i y_t(m-i) = - \sum_{i=o}^{s} \phi_i \sum_{j=1}^{m-i-1} \alpha_j e_{t+m-i-j}$$

$$+ \sum_{k=1}^{s} e_{t+m-k} \sum_{j=o}^{k-1} \phi_j \alpha_{k-j} + \sum_{k=s+1}^{r+s} e_{t+m-k} \sum_{j=o}^{s} \phi_j \alpha_{k-j}$$

$$= \sum_{k=m}^{r+s} e_{t+m-k} \beta_k \, .$$

Thus we have shown that for $1 \leq k \leq r+s$, if the k-step ahead predictor $y_t(k)$ satisfies conditions (i) and (ii) above, then it is identical to the k-step ahead predictor of an ARMA (s,r+s) process. Hence from equation (3.19), the predictors are identical for all $k \geq 1$.

Conversely, if $y_t(k)$ is the k-step ahead predictor of an ARMA (s,r+s) process, then there is a sequence $\alpha_1,\alpha_2,\ldots\alpha_{r+s}$ such that (3.20) holds (Box and Jenkins, 1970, P.134). Hence we can follow the argument of the first half of the proof to equation (3.27). Replacing t by t+1 in the expression for the ARMA model (2.5), we have

$$\sum_{i=o}^{s} \phi_i y_{t+1-i} = \sum_{i=o}^{r+s} \beta_i e_{t+1-i} .$$

e_{t+1-i} is the one step ahead prediction error $y_{t+1-i} - y_{t-i}(1)$; hence

$$\sum_{i=o}^{s} \phi_i y_{t-i}(1) = - \sum_{i=o}^{s} \phi_i e_{t+1-i} + \sum_{i=o}^{r+s} \beta_i e_{t+1-i}$$

$$= \sum_{i=o}^{s} (\beta_i-\phi_i) e_{t+1-i} + \sum_{i=s+1}^{r+s} \beta_i e_{t+1-i}. \quad (3.32)$$

Comparing (3.27) with (3.32), we find that

$$\phi_o = \beta_o = 1$$

$$\beta_i-\phi_i = \sum_{j=o}^{i-1} \phi_j \alpha_{i-j} \quad \text{or} \quad \beta_i = \sum_{j=o}^{i} \phi_j \alpha_{i-j} \qquad 1 \leq i \leq s$$

and

$$\beta_i = \sum_{j=o}^{s} \phi_j \alpha_{i-j} \qquad s+1 \leq i \leq r+s$$

which is (3.21) and (3.22).

Since we have an ARMA (s,r+s) model, the roots of $\beta(z)$ are greater than one in modulus, thus conditions (i) and (ii) are satisfied.

From the proof of this theorem, it can be seen that equivalences exist for other Box-Jenkins models. For example, if d of the roots of $\Phi(z)$ are unity, and the remainder lie outside the unit circle, then the following theorem can be proved.

Theorem 3.3

Let $r \geq 0$, $p > 0$, $d > 0$. Suppose the forecast function satisfies

$$\sum_{j=0}^{p+d} \phi_j y_t(k-j) = 0 \qquad k \geq p+d+r+1 \tag{3.33}$$

where the ϕ_i are real scalars with $\phi_0 = 1$ and

$$\Phi(z) = (1-z)^d \alpha(z)$$

where

$$\alpha(z) = 1+\alpha_1 z+\ldots+\alpha_p z^p$$

has all roots greater than one in modulus. It is assumed that p,d are the smallest integers for which such a representation exists and that (3.33) is not satisfied for $1 \leq k \leq p+d+r$. Then $y_t(k)$ is identical to the k-step ahead predictor of an ARIMA (p,d,p+d+r) process for each $k = 1,2,\ldots$ if and only if there is a real finite sequence $\lambda_0 = 1,\lambda_1,\ldots\lambda_{r+p+d}$ such that

(i) $y_t(k) = y_{t-1}(k+1) + \lambda_k e_t \qquad 1 \leq k \leq r+p+d$ (3.34)
where e_t is the one step ahead prediction error

(ii) the zeros of $\beta(z) = 1 + \beta_1 z+\ldots+\beta_{p+d+r} z^{p+d+r}$ lie strictly outside the unit circle, where β_j is given by (3.21) and (3.22).

Proof

This is proved in the manner of Theorem 3.2. In a similar manner, the proof of Theorem 3.2 will also provide conditions for equivalence to seasonal Box-Jenkins models.

CHAPTER 4

THE LINEAR MODEL

4.1 Introduction

The models considered in the previous two chapters are among the most popular for describing univariate time series. The extensive literature on this topic in both the theoretical and applied statistical journals is an indication of this. Many specialised applied journals, including the Journal of Accounting Research, Management Science, Journal of the Institute of Actuaries, Journal of Operations Research, IEEE proceedings and IEEE transactions on Automatic Control regularly contain papers on the use of ARIMA models in these fields.

However, it is widely accepted that there are difficulties with these models, the greatest of which seems to be the question of interpretation. In particular, the degrees of differencing and the autoregressive and moving average parameters have no intuitive interpretation. It is perhaps for this reason that attempts have been made to describe time series data in a way which is more intuitively acceptable. The linear model attempts to achieve this purpose. R.E. Kalman together with his co-workers (e.g. Kalman, 1960, Kalman and Bucy, 1961, Kalman, 1963a) is a major contributor to the extensive work on linear filters, much of which has accumulated in the control engineering literature. Some statistical papers e.g. those of Wishart (1969) and Whittle (1969), mention Kalman's work, and the contributions of Harrison (1967) and Harrison

and Stevens (1971, 1975, 1976) were also aimed primarily at statisticians. One of the achievements of these contributions is to give some meaning to the components of the time series. However, it turns out there may be some difficulty in selecting the appropriate dimension of the model, and this is bound up with the question of interpretation. Much of the remainder of this thesis will be concerned with the dimension of the models, primarily the Harrison-Stevens dynamic linear models (DLMs), their interpretation and their relationship to the ARIMA models described in Chapters 2 and 3.

We now consider some basic results on linear filters, and in particular on DLMs. The major sources of the material in this chapter and the next are Gelb (1974), Jacobs (1974), Sorenson (1966), Price (1974) and Harrison and Stevens (1976).

4.2 The Basic Model

We consider models of the form

$$\begin{aligned} \underline{Y}_t &= \underline{F}_t\underline{\theta}_t + \underline{v}_t \\ \underline{\theta}_t &= \underline{G}_t\underline{\theta}_{t-1} + \underline{H}_t\underline{w}_t \end{aligned} \qquad (4.1)$$

where $\underline{\theta}_t$ is a process vector varying in time, subject to the random term $\underline{H}_t\underline{w}_t$. The observations $\underline{Y}_t$ of the function $\underline{F}_t\underline{\theta}_t$ are made at discrete, not necessarily regular, intervals of time, and are subject to a random measurement error. The vectors $\underline{Y}_t, \underline{v}_t$ are of order $m \times 1$, $\underline{\theta}_t$ is of order $n \times 1$ and $\underline{w}_t$ is of order $r \times 1$. $\underline{F}_t, \underline{G}_t$ and $\underline{H}_t$ are matrices assumed known at time t, of dimension $m \times n$, $n \times n$ and $n \times r$ respectively. We

shall assume that the random vectors $\underline{v}_t$, $\underline{w}_t$ satisfy

$$E\left[\underline{w}_t\right] = \underline{0}, \qquad E\left[\underline{v}_t\right] = \underline{0},$$

$$E\left[\underline{v}_t\underline{v}_{t+k}^T\right] = \underline{V}_t\delta_{0,k} \qquad E\left[\underline{w}_t\underline{w}_{t+k}^T\right] = \underline{W}_t\delta_{0,k} \quad \text{for all } t,k$$

$$E\left[\underline{w}_t\underline{v}_{t+k}^T\right] = \underline{0} \quad \text{for all } t,k \tag{4.2}$$

In particular, if $\underline{G}_t$ is independent of t, and $\underline{H}_t = \underline{I}$, and with the additional assumption that $\underline{v}_t, \underline{w}_t$ are normally distributed, the model (4.1) is a DLM of the form given by Harrison and Stevens (1976). The assumption of normality is often made, and while it is not essential for the results of this chapter, it is often helpful in the interpretation of results.

In general, we are interested in $\underline{\theta}_t$ at time t, or in predicting, at time t, the values of $\underline{Y}_{t+k}$ or $\underline{\theta}_{t+k}(k \geq 1)$. These quantities have to be estimated from the available data $\underline{Y}^t = (\underline{Y}_1, \underline{Y}_2, \ldots \underline{Y}_t)$.

Given an estimate $\hat{\underline{\theta}}_{t-1}$ of $\underline{\theta}_{t-1}$ at time $t-1$, we can form an intuitive estimate of $\underline{\theta}_t$ by taking expectations of the 'system equation' of (4.1) to give

$$\underline{\theta}_t^* = \underline{G}_t\hat{\underline{\theta}}_{t-1}. \tag{4.3}$$

But this takes no account of the information in $\underline{Y}_t$. Given this estimate $\underline{\theta}_t^*$, we expect the current observation $\underline{Y}_t$ to have the value $\underline{F}_t\underline{\theta}_t^* = \underline{F}_t\underline{G}_t\hat{\underline{\theta}}_{t-1}$, so that the discrepancy between the observed and the expected observation is

$$\underline{Y}_t - \underline{F}_t\underline{G}_t\hat{\underline{\theta}}_{t-1}.$$

It seems intuitively obvious that the estimate of $\underline{\theta}_t$ should be modified in proportion to this discrepancy. Since we are only concerned with linear filters, the estimator is formed by the intuitive estimator (4.3) plus some

weighting matrix times this discrepancy. Expressed mathematically

$$\hat{\underline{\theta}}_t = \underline{G}_t\hat{\underline{\theta}}_{t-1} + \underline{A}_t(\underline{y}_t - \underline{F}_t\underline{G}_t\hat{\underline{\theta}}_{t-1}). \qquad (4.4)$$

Depending on what is required of the estimator, $\underline{A}_t$ may be a fixed constant matrix, or may vary in time according to some rule.

Example 4.1

A so-called constant velocity model used by engineers is given by

$$y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_t \\ \dot{x}_t \end{bmatrix} + v_t$$

$$\begin{bmatrix} x_t \\ \dot{x}_t \end{bmatrix} = \begin{bmatrix} 1 & \tau \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ \dot{x}_{t-1} \end{bmatrix} + \begin{bmatrix} \tau^2/2 \\ \tau \end{bmatrix} a_t ,$$

where τ is the interval between measurements, and a_t is random acceleration, $E(a_t) = 0$. When τ is constant, the estimators of $\begin{bmatrix} x_t & \dot{x}_t \end{bmatrix}^T$ are commonly found using the α-β filter, which is (4.4) with $\underline{A}_t$ given by $\begin{bmatrix} \alpha_t, & \beta_t/\tau \end{bmatrix}^T$. In many contexts, α_t and β_t are chosen as fixed constants throughout, usually between 0 and 1. If a least-squares fit through the first t data points is required, then the components of the $\underline{A}_t$ matrix are given by

$$\alpha_t = 2(2t-1)/t(t+1), \quad \beta_t = 6/t(t+1).$$

In this case, the filter is known as the expanding memory filter. If the random acceleration is thought to make a significant difference, we may choose to weight the observations, so that more attention is paid to the more recent data. In this case, a fading memory filter may be used. The detailed discussion of this and similar filters is given in Morrison (1969).

Kalman (1963a) suggested we should choose $\underline{A}_t$ to minimise $E\left[(\underline{\theta}_t - \hat{\underline{\theta}}_t)^T(\underline{\theta}_t - \hat{\underline{\theta}}_t)\right]$. We note that

$$E\left[(\underline{\theta}_t-\hat{\underline{\theta}}_t)^T(\underline{\theta}_t-\hat{\underline{\theta}}_t)\right] = \text{trace } E\left[(\underline{\theta}_t-\hat{\underline{\theta}}_t)(\underline{\theta}_t-\hat{\underline{\theta}}_t)^T\right]$$

and define

$$\underline{C}_t = E\left[(\underline{\theta}_t-\hat{\underline{\theta}}_t)(\underline{\theta}_t-\hat{\underline{\theta}}_t)^T\right]. \tag{4.5}$$

Assuming $\hat{\underline{\theta}}_t$ is an unbiased estimator of $\underline{\theta}_t$, $\underline{C}_t$ is the covariance matrix of $\underline{\theta}_t$. Further discussion on the properties of $\hat{\underline{\theta}}_t$ takes place later in this section.

Choosing $\underline{A}_t$ such that trace $(\underline{C}_t)$ is minimised yields

$$\underline{A}_t = \underline{P}_t\underline{F}_t^T(\underline{F}_t\underline{P}_t\underline{F}_t^T + \underline{V}_t)^{-1} \tag{4.6}$$

where

$$\underline{P}_t = \underline{G}_t\underline{C}_{t-1}\underline{G}_t^T + \underline{H}_t\underline{W}_t\underline{H}_t^T \tag{4.7}$$

and

$$\underline{C}_t = (\underline{I}-\underline{A}_t\underline{F}_t)\underline{P}_t \ . \tag{4.8}$$

Notice that $\underline{P}_t$ is the covariance matrix of the intuitive estimate (4.3).

If $\underline{V}_t$ is positive definite, then $\underline{A}_t$ can be expressed in a form more open to interpretation. First note that

$$\underline{C}_t^{-1} = \underline{P}_t^{-1} + \underline{F}_t^T\underline{V}_t^{-1}\underline{F}_t \ . \tag{4.9}$$

This is easily checked by multiplying together the expessions (4.8) and (4.9) for $\underline{C}_t$ and $\underline{C}_t^{-1}$. Premultiply the right hand side of equation (4.6) by $\underline{C}_t\underline{C}_t^{-1}$ to obtain

$$\begin{aligned}\underline{A}_t &= \underline{C}_t(\underline{P}_t^{-1}+\underline{F}_t^T\underline{V}_t^{-1}\underline{F}_t)\underline{P}_t\underline{F}_t^T(\underline{F}_t\underline{P}_t\underline{F}_t^T+\underline{V}_t)^{-1}\\ &= \underline{C}_t\underline{F}_t^T(\underline{I}+\underline{V}_t^{-1}\underline{F}_t\underline{P}_t\underline{F}_t^T)(\underline{F}_t\underline{P}_t\underline{F}_t^T+\underline{V}_t)^{-1}\\ &= \underline{C}_t\underline{F}_t^T\underline{V}_t^{-1} \ .\end{aligned} \tag{4.10}$$

$\underline{F}_t$ is merely a coefficient matrix, so suppose for simplicity that $\underline{F}_t = \underline{I}$ and $\underline{V}_t$ is diagonal. Then each element of $\underline{A}_t$ is proportional to the uncertainty of the estimate and inversely proportional to the measurement noise. It follows that if the measurement noise is large and estimation errors

are small, then $\underline{A}_t$ is small. Thus little attention is paid to the most recent observation because we have more confidence in the previous estimator. Conversely, if measurement noise is small and estimation errors are large, then $\underline{A}_t$ is large, demonstrating the need for more information.

From equations (4.4) - (4.8) we see that we need an initial estimate of $\underline{\theta}_o$, together with its initial covariance matrix $\underline{C}_o$. If we define

$$\hat{\underline{\theta}}_o = E\left[\underline{\theta}_o\right]$$

with

$$\underline{C}_o = E\left[(\underline{\theta}_o - \hat{\underline{\theta}}_o)(\underline{\theta}_o - \hat{\underline{\theta}}_o)^T\right],$$

then $\hat{\underline{\theta}}_t$ is an unbiased estimator of $\underline{\theta}_t$ for all t, and is by definition the minimum variance estimator. A proof of this result is given in Sorenson (1966). In practice, it is not always possible to define $\hat{\underline{\theta}}_o$ in this way. But it has been shown that for large t, the effects of the initialisation are negligible, so that $\hat{\underline{\theta}}_t$ can be regarded as an unbiased estimator of $\underline{\theta}_t$ for all sufficiently large t.

The Harrison-Stevens DLMs assume that $\underline{\theta}_o$ is normally distributed, and since $\underline{v}_t$, $\underline{w}_t$ are also normally distributed, it can be shown by Bayes theorem that all subsequent $\underline{\theta}_t|\underline{y}^t$, are normally distributed with mean $\hat{\underline{\theta}}_t$ and covariance matrix $\underline{C}_t$ as defined in equations (4.4) - (4.8). In this case, the Kalman updating procedure yields the minimum variance estimator of $\underline{\theta}_t$.

If the model is not a DLM, for example if any distribution is not normal, then the Kalman filter yields the minimum variance <u>linear</u> estimator of $\underline{\theta}_t$. In theory, the distribution of $\underline{\theta}_t|\underline{y}_t$ can be found from Bayes' theorem, but unless all

the distributions involved are normal, the analysis is likely to become quickly intractable. In general, we know only that the mean is $\hat{\underline{\theta}}_t$ and the covariance matrix is $\underline{C}_t$.

Example 4.2

The steady model given by Harrison and Stevens (1976) can be used whenever the 'level' of the process is more or less constant. It can be written as

$$y_t = u_t + \varepsilon_t$$

$$u_t = u_{t-1} + \delta u_t \ .$$

Here u_t is the true level of the process at time t, and y_t is the observed level. In DLM notation, we have $m = n = 1$, $F_t = G = 1$ for all t and $V_t = \text{var}(\varepsilon_t)$, $W_t = \text{var}(\delta u_t)$. Applying the Kalman filter equations (4.4) - (4.8) yields

$$\hat{u}_t = \hat{u}_{t-1} + A_t(y_t - \hat{u}_{t-1})$$

where

$$A_t = (C_{t-1}+\text{var}(\delta u_t))/(C_{t-1}+\text{var}(\delta u_t)+\text{var}(\varepsilon_t))$$

and

$$C_t = A_t \text{var}(\varepsilon_t) \ .$$

Two of the most often quoted uses of this model are for sales of an established product, where demand is almost constant, and for the position of a supposedly stationary object. It should however be noted that it is possible to describe this situation as a DLM with $\underline{G}$ having dimension greater than one. This is discussed in Chapter 6. It is also possible to introduce a 'slope' term, so that the 'level' is increasing (or decreasing) at a fairly constant rate. Intuitively, this involves two state variables and is often called a linear growth model. It

corresponds to a polynomial model of degree one, while the steady model corresponds to a polynomial of degree zero. It is possible to define polynomial models of degree d-1, and Harrison (1967) has pointed out that intuitively, these require d state variables. It is argued here that more than d state variables are preferable to describe a polynomial of degree d-1. The reasons for this are discussed in Chapter 7. Forward shifted polynomial models, where the model behaves like a polynomial model only after a certain lag has passed, can also be defined. These are described in Chapter 8. More generally, we can define DLMs which are asymptotically polynomial models, and others which are seasonal in nature.

4.3 Forecasting DLMs

In this section, we shall derive the forecasts $\underline{Y}_t(k)$ of $\underline{Y}_{t+k}$ at time t for the DLM. From (4.1) and the following discussion, the DLM at time $t+k$ is given by

$$\begin{aligned} \underline{Y}_{t+k} &= \underline{F}_{t+k}\underline{\theta}_{t+k} + \underline{v}_{t+k} \\ \underline{\theta}_{t+k} &= \underline{G}\underline{\theta}_{t+k-1} + \underline{w}_{t+k} \cdot \end{aligned} \tag{4.11}$$

Taking expectations of $\underline{\theta}_{t+k}$ conditional on $\underline{Y}^t$, we have

$$E\left[\underline{\theta}_{t+k}|\underline{Y}^t\right] = \underline{G}\, E\left[\underline{\theta}_{t+k-1}|\underline{Y}^t\right]$$

with associated covariance matrix

$$\mathrm{cov}\left[\underline{\theta}_{t+k}|\underline{Y}^t\right] = \underline{G}\, \mathrm{cov}\left[\underline{\theta}_{t+k-1}|\underline{Y}^t\right]\underline{G}^T + \underline{W}_{t+k} \cdot$$

Thus the mean and covariance matrix of $\underline{\theta}_{t+k}$ can be calculated recursively, using $E\left[\underline{\theta}_t|\underline{Y}^t\right] = \hat{\underline{\theta}}_t$ as defined in equation (4.4). By induction

$$E\left[\underline{\theta}_{t+k}|\underline{Y}^t\right] = \underline{G}^k\hat{\underline{\theta}}_t$$

with covariance matrix

$$\text{cov}\left[\underline{\theta}_{t+k}|\underline{Y}^t\right] = \underline{G}^k\underline{C}_t(\underline{G}^k)^T + \sum_{i=0}^{k-1} \underline{G}^i\underline{W}_{t+k-i}(\underline{G}^i)^T \ .$$

Moving on to the statistically more interesting problem of prediction of $\underline{Y}_{t+k}$ at time t,

$$E\left[\underline{Y}_{t+k}|\underline{Y}^t\right] = \underline{F}_{t+k}E\left[\underline{\theta}_{t+k}|\underline{Y}^t\right] = \underline{F}_{t+k}\underline{G}^k\hat{\underline{\theta}}_t \qquad (4.12)$$

and

$$\text{cov}\left[\underline{Y}_{t+k}|\underline{Y}^t\right] = \underline{F}_{t+k}\text{cov}\left[\underline{\theta}_{t+k}|\underline{Y}^t\right]\underline{F}_{t+k}^T + \underline{V}_{t+k}$$

where it is assumed that $\underline{F}_{t+k}$ is known at time t. In fact, in many applications, and for the major part of this thesis, $\underline{F}_t$ will be a constant matrix, usually a row vector.

Having obtained the mean and covariance matrix of $\underline{Y}_{t+k}|\underline{Y}^t$, we now consider what to take as the predictor $\underline{Y}_t(k)$ of $\underline{Y}_{t+k}$. The loss function is defined as the real-valued function $L(\underline{Y}_{t+k}, \underline{Y}_t(k))$ which represents the cost if $\underline{Y}_t(k)$ is not an accurate estimate of $\underline{Y}_{t+k}$. It must therefore satisfy

$$\begin{aligned} L(\underline{Y}_{t+k}, \underline{Y}_t(k)) &= 0 \qquad \underline{Y}_t(k) = \underline{Y}_{t+k} \\ &\geq 0 \qquad \underline{Y}_t(k) \neq \underline{Y}_{t+k} \ . \end{aligned}$$

It is required to choose the predictor $\underline{Y}_t(k)$ so as to minimise the expected loss.

Although quadratic loss functions have been criticised as being unsuitable for many practical purposes, they are perhaps the best compromise in the absence of further information. Assuming a quadratic loss function, the expected loss is

$$\int_y (\underline{Y}_{t+k}-\underline{Y}_t(k))^T(\underline{Y}_{t+k}-\underline{Y}_t(k))f(\underline{Y}_{t+k}|\underline{Y}^t)d\underline{Y}_{t+k}$$

$$= \text{cov}(\underline{y}_{t+k} - E\left[\underline{y}_{t+k}|\underline{y}^t\right]) + (E\left[\underline{y}_{t+k}|\underline{y}^t\right] - \underline{y}_t(k))^T (E\left[\underline{y}_{t+k}|\underline{y}^t\right] - \underline{y}_t(k))$$

where $f(\underline{y}_{t+k}|\underline{y}^t)$ is the probability density function of $\underline{y}_{t+k}|\underline{y}^t$. The expected loss is clearly minimised when

$$\underline{y}_t(k) = E\left[\underline{y}_{t+k}|\underline{y}^t\right] . \tag{4.13}$$

This result is valid whatever the distribution of $\underline{y}_{t+k}|\underline{y}^t$.

Another result due to Sherman (1955) shows that if

(i) the loss function is symmetric about y_{t+k} and non-decreasing in $|y_{t+k} - y_t(k)|$

(ii) the distribution of $y_{t+k}|\underline{y}^t$ is symmetric about the mean and unimodal

then the expectation of the loss function is minimised when the predictor is given by the conditional expectation (4.13). Condition (ii) is satisfied by some of the most frequently used distributions, in particular the normal distribution is of this nature.

These two results both indicate that a sensible value for the predictor $\underline{y}_t(k)$ is the conditional expectation of $\underline{y}_{t+k}$, thus, $\underline{y}_t(k)$ will be defined by

$$\underline{y}_t(k) = \underline{FG}^k \hat{\underline{\theta}}_t . \tag{4.14}$$

Premultiplying the Kalman updating equation (4.4) by $\underline{FG}^k$, we have

$$\underline{FG}^k \hat{\underline{\theta}}_t = \underline{FG}^{k+1} \hat{\underline{\theta}}_{t-1} + \underline{FG}^k \underline{A}_t (\underline{y}_t - \underline{FG}\hat{\underline{\theta}}_{t-1})$$

or from (4.14)

$$\underline{y}_t(k) = \underline{y}_{t-1}(k+1) + \alpha_t (\underline{y}_t - \underline{y}_{t-1}(1)) . \tag{4.15}$$

This will be referred to as the predictor updating equation.

Example 4.3

The predictor of the Harrison-Stevens steady model is $y_t(k) = \hat{u}_t$ for all k, where $\hat{u}_t$ is as defined in Example 4.2. The variance, of course, increases as k increases:

$$E\left[(y_{t+k}-\hat{u}_t)^2\right] = C_t + \sum_{i=1}^{k} \text{var}(\delta\mu_{t+i}) + \text{var}(\varepsilon_{t+k}).$$

Thus the predictor $y_t(k)$ is a constant, or in other words, the forecast function $\{y_t(k);\ k=1,2,\ldots\}$ is a polynomial of degree zero.

Similarly the Harrison-Stevens linear growth model has a forecast function

$$y_t(k) = a + bk$$

for all k, where a,b are quantities which are independent of k, but derived in terms of y^t, with $b \neq 0$, so it projects a straight line of non-zero gradient. By analogy, the polynomial projecting models of degree $d-1$ have forecast functions which are themselves polynomials of degree $d-1$, i.e.

$$y_t(k) = a_o + a_1 k + \ldots + a_{d-1}k^{d-1} \qquad k \geq 1$$

where the a_i are functions of y^t and independent of k, and $a_{d-1} \neq 0$. There are also models whose first r predictors seem to follow no particular pattern, but for $k > r$, the predictors form a polynomial of degree $d-1$. It is convenient to call these models forward shifted polynomial models of degree $d-1$ with shift r. The forecasts satisfy

$$y_t(k) = a_o + a_1 k + \ldots + a_{d-1}k^{d-1} \qquad k \geq r+1 \qquad (4.16)$$

but (4.16) is not satisfied for $1 \leq k \leq r$.

The above discussion has assumed univariate observations, but clearly could apply equally well to the elements of a vector of observations.

Smith (1979) considered the simplest form of Harrison-Stevens model, the steady model of Examples 4.2 and 4.3, when the normality assumption is relaxed. However, Key and Godolphin (1981) point out that his approach can lead to difficulties. In particular, the Kalman filter predictor updating equation (4.15) may not hold, and the forecasts do not necessarily behave in the manner expected of a 'steady model'.

4.4 Fitting DLMs to Time Series Data

It will be clear from Sections 4.2 and 4.3 that with a model of the form (4.1), forecasts can be made when there is little or even no data, using subjective values of $\underline{\theta}_0$ and $\underline{C}_0$. This is most useful: for example, if a new product is being launched, forecasts of demand will be needed at the beginning of the project. Similarly, in tracking aircraft, the position and velocity need to be roughly known immediately for control purposes.

Subjective information can be used at any time to modify the parameters of the system, not just for the initial estimates. For suppose it is thought, perhaps as the result of an advertising campaign, that $\underline{\theta}_{t+1}$ will be increased by an amount $\underline{\Psi}_{t+1}$ more than would usually be expected. Then $E\left[\underline{\theta}_{t+1}|\underline{y}^t\right]$ produced by the Kalman updating equation can be overridden, and increased by $\underline{\Psi}_{t+1}$, together with an increase in the covariance matrix to reflect the increased uncertainty. In this way, $\underline{A}_t$ is increased, giving extra weight to the next few

observations. The estimates should soon settle around the new true mean, and the effects of the increase in variance will soon fade away, but without this interference, the filter would take much longer to adjust for the increase and may not adjust completely for some time if $\underline{C}_t$ remains small.

Evidently, this type of information is more easily incorporated if the model is formed in a meaningful way, so that any subjective information can be expressed in terms of its effects on the state variables and observations.

In much of the literature, it is stressed that the matrices $\underline{F}$, $\underline{G}$, $\underline{V}$ and $\underline{W}$ need to be specified with care. For example, Jacobs (1974, p.310) states that the Kalman filter has been found to be sensitive to these matrices. The use of inaccurate matrices can cause the filter to diverge or 'learn the wrong state too well', a quotation from Jazwinski (1970, §8.8), where this form of divergence is discussed. The coefficient matrices are often known from the physical situation (at least to a linear approximation), but the specification of variances and covariances would be more difficult. Harrison and Stevens (1976) suggest that in commercial situations, these quantities usually have to be specified by non-statisticians. Although it may be possible for them to suggest the presence of correlation, it seems unlikely that they will be able to quantify the covariances. Because of the sensitivity of the Kalman filter to these quantities, it seems advisable to try to choose a model for which a diagonal system covariance matrix can be specified.

This may not be very difficult in practice, for more than one interpretation of the system may be available,

and it may be that one of these has a diagonal covariance matrix. Indeed, in theory, every model has a DLM representation for which a diagonal system covariance matrix is applicable. For example, consider the model

$$y_t = 2\theta_t - \psi_t + \varepsilon_t$$

$$\theta_t = \psi_{t-1} + w_{1t}$$

$$\psi_t = -\theta_{t-1} + 2\psi_{t-1} + w_{2t}$$

where

$$\underline{W}_1 = \begin{bmatrix} \mathrm{var}(w_{1t}) & \mathrm{cov}(w_{1t},w_{2t}) \\ \mathrm{cov}(w_{1t},w_{2t}) & \mathrm{var}(w_{2t}) \end{bmatrix}$$

is independent of time.

This model satisfies

$$\nabla^2 y_t = 2w_{1t} - 3w_{1t-1} - w_{2t} + 2w_{2t-1} + \nabla^2\varepsilon_t \qquad (4.17)$$

and

$$E\left[(\nabla^2 y_t)^2\right] = 13\mathrm{var}(w_{1t}) - 16\mathrm{cov}(w_{1t},w_{2t}) + 5\mathrm{var}(w_{2t}) + 6\mathrm{var}(\varepsilon_t)$$

$$E\left[(\nabla^2 y_t)(\nabla^2 y_{t-1})\right] = -6\mathrm{var}(w_{1t}) - 2\mathrm{var}(w_{2t}) + 7\mathrm{cov}(w_{1t},w_{2t}) - 4\mathrm{var}(\varepsilon_t)$$

$$E\left[(\nabla^2 y_t)(\nabla^2 y_{t-2})\right] = \mathrm{var}(\varepsilon_t) \quad .$$

At first sight, this looks very different from the linear growth model described by Harrison and Stevens (1976), which will be discussed in more detail in Chapter 7. This model has

$$\underline{W}_2 = \begin{bmatrix} \mathrm{var}(\delta u_t) & 0 \\ 0 & \mathrm{var}(\delta\beta_t) \end{bmatrix}$$

and satisfies

$$\nabla^2 y_t = \delta\beta_t + \nabla\delta u_t + \nabla^2\varepsilon_t \qquad (4.18)$$

and

$$E\left[(\nabla^2 y_t)^2\right] = \text{var}(\delta\beta_t) + 2\text{var}(\delta u_t) + 6\text{var}(\varepsilon_t)$$

$$E\left[(\nabla^2 y_t)(\nabla^2 y_{t-1})\right] = -\text{var}(\delta u_t) - 4\text{var}(\varepsilon_t)$$

$$E\left[(\nabla^2 y_t)(\nabla^2 y_{t-2})\right] = \text{var}(\varepsilon_t) \ .$$

However, there is at least one set of values $\text{var}(w_{1t})$, $\text{var}(w_{2t})$, $\text{cov}(w_{1t},w_{2t})$ for each set $\text{var}(\delta\beta_t)$, $\text{var}(\delta u_t)$ such that the two models have identical covariance structure and hence are different representations of the same model. For example, let $\text{var}(\delta\beta_t)$, $\text{var}(\delta u_t)$ be given and suppose w_{1t}, w_{2t} are such that

$$\text{var}(w_{1t}) = \text{var}(\delta u_t) + \text{var}(\delta\beta_t)$$

$$\text{var}(w_{2t}) = \text{var}(\delta u_t) + 4\text{var}(\delta\beta_t)$$

$$\text{cov}(w_{1t},w_{2t}) = \text{var}(\delta u_t) + 2\text{var}(\delta\beta_t).$$

Then we see that every linear growth model satisfying (4.18), where $\text{cov}(\delta u_t,\delta\beta_t) = 0$, can be written as the above example which leads to (4.17). The converse is not true, for $\text{cov}(w_{1t},w_{2t})$ must be strictly positive.

Different representations of this sort are possible because the behaviour of the system is not determined by $\underline{F}$ and $\underline{G}$ themselves, but by some of their properties. The important properties, listed by Jacobs (1974) for deterministic models, and shown here to be applicable to stochastic models are:

(a) the number of state variables, n

(b) the eigenvalues and eigenvectors of $\underline{G}$

(c) the linear dependence/independence of $\underline{F}$, $\underline{FG}$,...$\underline{FG}^{n-1}$

(d) for stochastic models, the rank of $\underline{W}$.

All the above properties are invariant under linear invertible transformations of the state vector $\underline{\theta}_t$. For suppose

$$\underline{x}_t = \underline{L}\underline{\theta}_t$$

where $\underline{L}$ is non-singular. Then

$$\underline{Y}_t = \underline{F}'\underline{x}_t + \underline{v}_t$$

$$\underline{x}_t = \underline{G}'\underline{x}_{t-1} + \underline{w}'_t$$

where

$$\underline{F}' = \underline{F}\underline{L}^{-1}, \ \underline{G}' = \underline{L}\underline{G}\underline{L}^{-1} \text{ and } \underline{w}'_t = \underline{L}\underline{w}_t.$$

Thus $\underline{F}'(G')^k = \underline{F}\underline{G}^k\underline{L}^{-1}$. The Harrison-Stevens linear growth model was formed from the first by

$$\underline{L} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}.$$

Since any positive definite $\underline{W}$ can be diagonalised by an appropriate non-singular transformation $\underline{L}\underline{W}\underline{L}^T$, every model has a DLM representation with diagonal $\underline{W}$. There are two problems.

1. It is possible that this representation has no meaningful interpretation, making it difficult to use subjective information.
2. The matrix $\underline{L}$ may involve the covariances, so the above method may fail to be directly applicable in practice.

Section 4.1 of Harrison and Stevens (1976) gives an indication of a possible method of estimation of $\underline{V}$ and $\underline{W}$ using several values together with their probabilities and updating their probabilities at each stage.

In their paper, Harrison and Stevens verify that a linear combination of models is itself a linear model.

Thus a complex system can be broken down into simple DLMs and then combined into a whole. The obvious application is a simple growth process with seasonal effects. In this case, the growth process and the seasonal process can be modelled separately and then combined into one model.

Suppose the two models are given by

$$\underline{Y}_{it} = \underline{F}_i \underline{\theta}_{it}$$

$$\underline{\theta}_{it} = \underline{G}_{ii} \underline{\theta}_{it-1} + \underline{w}_{it} . \qquad i = 1,2$$

Then the complete model is

$$\underline{Y}_t = \underline{Y}_{1t} + \underline{Y}_{2t} + \underline{v}_t = \begin{bmatrix} \underline{F}_1 & \underline{F}_2 \end{bmatrix} \begin{bmatrix} \underline{\theta}_{1t} \\ \underline{\theta}_{2t} \end{bmatrix} + \underline{v}_t$$

$$\begin{bmatrix} \underline{\theta}_{1t} \\ \underline{\theta}_{2t} \end{bmatrix} = \begin{bmatrix} \underline{G}_{11} & \underline{0} \\ \underline{0} & \underline{G}_{22} \end{bmatrix} \begin{bmatrix} \underline{\theta}_{1t-1} \\ \underline{\theta}_{2t-1} \end{bmatrix} + \begin{bmatrix} \underline{w}_{1t} \\ \underline{w}_{2t} \end{bmatrix} .$$

If there is interaction between $\underline{\theta}_{1t}$ and $\underline{\theta}_{2t}$, then non-zero matrices $\underline{G}_{12}$, $\underline{G}_{21}$ could be introduced, or non-zero covariances between the elements of the system noise vector. This leads us back to the problem of how to specify the covariance matrices $\underline{V}$ and $\underline{W}$.

4.5 Comparison of Linear Control Theory Models and Classical Time Series Models

In this thesis we shall be concerned with the comparison of the predictors of the linear control theory model with those of the univariate time series models considered in the previous two chapters. Thus in this section and in Chapters 6-9, we shall assume $m = 1$, that is y_t, v_t are scalars. The coefficients of the

time series models do not change with time, so if the linear models given by (4.1) are to be comparable, then $\underline{F},\underline{G},\underline{H},\underline{V}$ and $\underline{W}$ must be independent of time. For simplicity, and without significant loss of generality, we shall assume $H = I$, thus (4.1) becomes

$$\begin{aligned} Y_t &= \underline{F}\underline{\theta}_t + v_t \\ \underline{\theta}_t &= \underline{G}\underline{\theta}_{t-1} + \underline{w}_t . \end{aligned} \tag{4.19}$$

It has been argued by Akaike (1974a) that the set of models given by (4.19) and the set of ARMA models given by (2.5) are equivalent, subject to certain conditions which we discuss here.

Akaike considers the restricted model

$$Y_t = \underline{F}\underline{\theta}_t \tag{4.20}$$

$$\underline{\theta}_t = \underline{G}\underline{\theta}_{t-1} + \underline{w}_t \tag{4.21}$$

i.e.(4.19) with the observation noise absent. The characteristic polynomial of $\underline{G}$ is

$$\det(\lambda\underline{I} - \underline{G}) = \sum_{k=0}^{n} a_k \lambda^{n-k} \tag{4.22}$$

with $a_0 = 1$. Then by the Cayley-Hamilton theorem that every square matrix satisfies its own characteristic equation

$$\underline{G}^n + \sum_{k=1}^{n} a_k \underline{G}^{n-k} = \underline{O} . \tag{4.23}$$

Also, substituting (4.21) successively into itself we obtain for $0 \leq k \leq n-1$

$$\underline{\theta}_{t-k} = \underline{G}^{n-k}\underline{\theta}_{t-n} + \sum_{i=0}^{n-k-1} \underline{G}^i \underline{w}_{t-k-i} . \tag{4.24}$$

Therefore using equation (4.20) and $a_1,\ldots,a_n$ defined by equation (4.22)

$$Y_t + \sum_{i=1}^{n} a_i Y_{t-i} = \underline{F}\left[\underline{\theta}_t + \sum_{i=1}^{n} a_i \underline{\theta}_{t-i}\right] .$$

Substituting for $\underline{\theta}_{t-k}$, $0 \leq k \leq n-1$ from equation (4.24), the right hand side becomes

$$\underline{F} \sum_{i=0}^{n-1} a_i \left[\sum_{j=0}^{n-i-1} \underline{G}^j \underline{w}_{t-j-i} + \underline{G}^{n-i} \underline{\theta}_{t-n} \right] + a_n \underline{F} \underline{\theta}_{t-n}$$

$$= \underline{F} \left[\sum_{k=0}^{n-1} \underline{w}_{t-k} \sum_{i=0}^{k} a_i \underline{G}^{k-i} + \sum_{i=0}^{n} a_i \underline{G}^{n-i} \underline{\theta}_{t-n} \right]$$

so that from equation (4.23)

$$y_t + \sum_{i=1}^{n} a_i y_{t-i} = \underline{F} \underline{w}_t + \underline{F}_1 \underline{w}_{t-1} + \ldots + \underline{F}_{n-1} \underline{w}_{t+1-n} \qquad (4.25)$$

where

$$\underline{F}_i = \left[\underline{F} \underline{G}^i + a_1 \underline{G}^{i-1} + \ldots + a_i \underline{I} \right]. \qquad (4.26)$$

Equation (4.25) provides a representation for the univariate time series $\{y_t\}$ which can be expressed as an ARMA process under restrictive conditions on (4.20) and (4.21). For example, if $\underline{F} = [1\ 0\ 0 \ldots 0]$ and $\underline{G}$ is lower triangular, then the right hand side of (4.25) can be expressed as a moving average of order n-1. Other representations for $\underline{F}, \underline{G}, \underline{W}$ which yield a scalar moving average for the right hand side of (4.25) exist, but it seems these need to be rather specialised and interdependent. Further restrictions on (4.20) and (4.21) are required if (4.25) is to be stationary and invertible, but this point is not discussed in detail by Akaike. However Akaike does give a converse argument showing that an ARMA model (2.5) can be written in the form (4.20) and (4.21) but this argument relies upon the Wold decomposition theorem, hence it does not apply to the interesting non-stationary ARIMA models, which we wish to consider.

Furthermore, our main interest lies not in the various models which could be employed, but in the observations

themselves and in the prediction of future observations. Demonstrating that two different forms of model are equivalent does not necessarily mean that the predictors will be identical because this will also depend on the prediction methods used for the model. For example, the ARMA models use a minimum mean square error prediction criterion, but the dynamic linear models can give a variety of different forecasts if subjective information is employed consistently. We adopt the conditional expectations, based on a minimum variance estimate of $\underline{\theta}_t$, i.e. $E\left[\underline{\theta}_{t+k} | \hat{\underline{\theta}}_t\right]$, where $\hat{\theta}_t$ is given by (4.4), as a forecasting criterion for dynamic linear models. The emphasis will be placed on demonstrating equivalence of predictors, not models.

CHAPTER 5

SOME CONTROL THEORY TOOLS AND CONCEPTS

5.1 z-transform theory and stability

We now consider a tool which will be frequently used in this work, namely the z-transform. This useful device converts linear difference equations such as equation (4.4) into algebraic equations, which are much easier to manipulate. It has many uses other than those to be discussed here, for example, Ray and Wyld (1965) used z-transform theory when investigating the presence of bias in certain polynomial projecting predictors. This topic is also discussed in Moon (1977), where the bias is called systematic error.

Definition

Let $\{x_t\}$ be a sequence defined for $t = 0,1,2,\ldots$. Then the z-transform $X(z)$ of x_t is defined as

$$X(z) = \sum_{t=0}^{\infty} x_t z^{-t} \qquad (5.1)$$

where z is a complex variable. It will sometimes be helpful to use the notation $z(x_t)$ to denote $X(z)$.

The z-transform is effectively the discrete version of the Laplace transform, in fact it can be derived as the Laplace transform of a continuous process x_t^* sampled at discrete intervals T to give a sequence $\{x_t\}$. This can be written mathematically as

$$x_t = \sum_{i=0}^{\infty} x_t^* \delta(t-iT)$$

where $\delta(t)$ is the Dirac delta function. Taking Laplace transforms of the discrete process x_t, we have

$$\mathcal{L}\left[x_t\right] = \int_0^\infty \sum_{k=0}^{\infty} x_t^* \delta(t-kT) e^{-st} dt$$

$$= \sum_{k=0}^{\infty} \int_0^\infty x_t^* \delta(t-kT) e^{-st} dt$$

$$= \sum_{k=0}^{\infty} x_{kT} e^{-ksT}.$$

Writing $z = e^{sT}$, we obtain

$$\mathcal{L}\left[x_t\right] = \sum_{k=0}^{\infty} x_{kT} z^{-k} = X(z) \quad \text{as required.}$$

In many books, for example, Oppenheim and Schafer (1975, Chapter 2), equation (5.1) defines the one sided z-transform, while the two sided z-transform is given by

$$X(z) = \sum_{t=-\infty}^{\infty} x_t z^{-t} . \tag{5.2}$$

Obviously, if $x_t = 0$ for all $t < 0$, then the two forms are identical, but not otherwise. In the applications used in this thesis, all the filters are physically realizable, that is $x_t = 0$ for all $t < 0$, so that the one- and two-sided z-transforms are identical. Since z-transforms are often written as a quotient of two polynomials, we need a test of physical realizability. The criterion we apply is given in Lindorff (1965) and can be written

$$\lim_{z \to \infty} z^{-1} X(z) = 0 \tag{5.3}$$

if and only if the filter is physically realizable. This criterion is easily seen to be valid because (5.2) reduces to (5.1) when (5.3) holds.

Properties of z-transforms

1. Linearity

$$z(\alpha x_t+\beta y_t) = \alpha z(x_t) + \beta z(y_t)$$

$$= \alpha X(z) + \beta Y(z) \qquad (5.4)$$

where α,β are arbitrary constants.

2. Shift of sequence

$$z(x_{t+a}) = z^a X(z) \quad \text{for} \quad a < 0 . \qquad (5.5)$$

3. Multiplication by an exponential sequence

$$z(a^t x_t) = X(z/a) . \qquad (5.6)$$

4. Convolution of sequences

Let $x_t = \sum_{k=0}^{\infty} \beta_k y_{t-k}$ where $\beta_i, y_i = 0$ for $i < 0$.

Then $X(z) = \beta(z)Y(z)$. (5.7)

5. Initial Value Theorem

If $X(z)$ denotes the z-transform of x_t, then the initial value x_t is given by

$$x_o = \lim_{z\to\infty} X(z) .$$

6. Final Value Theorem

If $X(z)$ denotes the z-transform of x_t, and $(z-1)X(z)$ has no poles on or outside the unit circle, then

$$\lim_{k\to\infty} x_{kT} = \lim_{z\to 1} (z-1)X(z) .$$

The first five properties follow directly from the defining equation (5.1). However, the sixth requires a little more subtlety, and a proof is given in Lindorff (1965, p.51).

From equation (5.5), it is easily shown that the z-transform of ∇X_t is $(1-z^{-1})X(z)$. By repeated application, we find that $z(\nabla^k X_t) = (1-z^{-1})^k X(z)$.

Suppose the output x_t of a filter is related to the input y_t by a difference equation

$$\sum_{k=0}^{p} \alpha_k x_{t-k} = \sum_{k=0}^{q} \beta_k y_{t-k} \qquad (5.8)$$

where $\{\alpha_k\}$, $\{\beta_k\}$ form known sequences, and p,q are any positive integers. The form of equation (5.8) is similar to that of the ARIMA process (2.5), but this similarity is only apparent, because the sequences $\{y_t\}$, $\{x_t\}$ in equation (5.8) are meant to denote any time sequence while the sequences $\{\varepsilon_t\}$ and $\{y_t\}$ of equation (2.5) have a specific structure. Let the z-transforms of α_k, β_k, x_k, y_k be given by $\alpha(z)$, $\beta(z)$, $X(z)$ and $Y(z)$ respectively. Then using property (4) of z-transforms, we find that

$$\alpha(z)X(z) = \beta(z)Y(z)$$

$$X(z) = \frac{\beta(z)}{\alpha(z)} Y(z). \qquad (5.9)$$

The ratio $G(z) = \beta(z)/\alpha(z)$ is known as the transfer function of the filter, and the denominator $\alpha(z)$ is known as the characteristic polynomial. It sometimes happens that the sequence $\{y_t\}$ and the transfer function $G(z)$ are known, hence the z-transform of the output sequence $X(z)$ is known, and it is required to find the output sequence $\{x_k\}$ (or some part of it) from $X(z)$. This is done by means of the inversion integral, which is given by

$$x_{kT} = \frac{1}{2\pi i} \oint_c X(z) z^{k-1} dz \qquad (5.10)$$

where c is a closed curve enclosing all the poles of $X(z)z^{k-1}$. This can also be expressed as

$$x_{kT} = \sum_{\text{poles}} \text{residue} \left[X(z) z^{k-1} \right].$$

One useful application of the transfer function is to determine whether or not the filter is stable, in the sense defined below.

Definition

A discrete linear filter is defined as stable if and only if the output is bounded in response to every bounded input (Lindorff 1965).

Expressed in terms of the filter defined in equation (5.8), when all the y_t are finite, then if the filter is stable it follows that all the x_t are necessarily finite. This appears to be a most important property for a filter to have, for it seems unlikely that a process which becomes infinite can be useful in practice. The following criterion for stability of a discrete linear filter can be found in Lindorff (1965, p.43).

Theorem 5.1

Given a linear discrete physically realizable filter with transfer function $G(z)$, a necessary and sufficient criterion for stability is $\sum_{t=0}^{\infty} |g_t| < \infty$.

An equivalent criterion for stability can be obtained from Theorem 5.1 as follows:

Theorem 5.2

A linear discrete physically realizable filter is stable if and only if its transfer function $G(z)$ contains no poles on or outside the unit circle.

This criterion will be used throughout this work. The proof can be found in Lindorff (1965, p.44). An alternative formulation is that the characteristic polynomial has all roots inside the unit circle.

Comparing equation (5.8) with the ARIMA model (2.5) and using Theorem 5.2 on equation (5.9), we see that stability of the model in control theory is equivalent to stationarity in time series analysis, since the roots of the characteristic polynomial $\alpha(z) = \sum_{i=0}^{p} \alpha_i z^{-i}$ which must be less than one in modulus, are the inverse of the roots of $\sum_{i=0}^{p} \alpha_i z^{i}$, which must be greater than one in modulus for stationarity.

The z-transform of a vector or matrix can be defined by taking z-transforms of each of the elements. Suppose

$$\sum_{k=0}^{p} \underline{A}_k \underline{Y}_{t-k} = \sum_{k=0}^{q} \underline{B}_k \underline{w}_{t-k}$$

where $\{\underline{A}_k\}$, $\{\underline{B}_k\}$ are known sequences of matrices, $\underline{A}_k$ is $m \times m$, $\underline{B}_k$ is $m \times n$, $\underline{Y}_t$, $\underline{w}_t$ are vectors of dimension m, n respectively. Applying z-transforms

$$\underline{A}(z)\underline{Y}(z) = \underline{B}(z)\underline{Y}(z)$$

or

$$\underline{Y}(z) = (\underline{A}(z))^{-1}\underline{B}(z)\underline{W}(z) \qquad (5.11)$$

so that the transfer function is

$$\underline{G}(z) = (\underline{A}(z))^{-1}\underline{B}(z).$$

It is clear that the denominator of the transfer function is the same for all elements of the matrix (i.e. the determinant of $\underline{A}(z)$), hence we only need to find one element of the transfer function to determine whether or not the whole filter is stable.

The z-transform theory described in this section does not enable us to apply z-transforms to equation (4.4) directly, since this involves the z-transform of $\underline{A}_t\underline{y}_t$. This can be found, using the complex convolution theorem, as in Oppenheim and Schafer (1975, §2.3.9), but this is somewhat complicated. It turns out that we need only consider equation (4.4) in the steady state when $\underline{A}_t = \underline{A}$ for all t, and this simplifies the analysis considerably. Thus assuming the equilibrium state, the z-transform equation corresponding to (4.4) is

$$\underline{\theta}(z) = z^{-1}\underline{G\theta}(z) + \underline{A}(\underline{Y}(z) - z^{-1}\underline{FG\theta}(z)).$$

Consequently

$$\underline{\theta}(z) = z(z\underline{I} - (\underline{I}-\underline{AF})\underline{G})^{-1}\underline{AY}(z), \tag{5.12}$$

so for stability, all eigenvalues of $(\underline{I}-\underline{AF})\underline{G}$ must be less than one in modulus.

In many applications, the transfer function is the quotient of two polynomials as in equation (5.9), (5.11) and (5.12). From Theorem 5.2, if the model (as in 5.9) or filter (as in 5.12) is to be stable, then the transfer function should have no poles on or outside the unit circle. Therefore, we have to determine whether or not the roots $z_1, z_2, \ldots z_r$ of a polynomial

$$\begin{aligned}\Psi(z) &= \psi_o + \psi_1 z + \psi_2 z^2 + \ldots + \psi_r z^r \\ &= \psi_r(z-z_1)\ldots(z-z_r)\end{aligned} \tag{5.13}$$

are less than one in modulus. Thus in (5.9), $\Psi(z)$ will be $\alpha(z)$, with $r = p$, while in (5.12) $\Psi(z)$ is the determinant of $(z\underline{I} - (\underline{I}-\underline{AF})\underline{G})$ and r is the dimension of $\underline{\theta}(z)$, i.e. $r = n$. This problem and its continuous time equivalent, that the roots of a polynomial should have negative real parts, have occupied many workers in several fields for over a century. These include Routh (1905), Schur (1917), Cohn (1922), Wold (1938), Wise (1956), Gantmacher (1959), Astrom (1970), Pagano (1973) and Anderson (1975b,1977). The text of Jury (1964) provides possibly the most comprehensive solution to the problem.

5.2 Jury's Stability Criterion

Jury gives three methods for testing whether or not the roots of a given polynomial $\Psi(z)$ lie within the unit circle. These methods can be described as the determinant method, the table method and the division method. The determinant method seems to be the most applicable for small or moderately large r, while the table method seems appropriate for numerical work, because it requires evaluation of only second order determinants, and is easily programmable on a digitial computer.

Jury's Determinant Method

To discuss this method, we first need to define the two matrices

$$\underline{X}_k = \begin{bmatrix} \psi_r & \psi_{r-1} & \psi_{r-2} & \cdots & \psi_{r-k+1} \\ 0 & \psi_r & \psi_{r-1} & & \psi_{r-k+2} \\ 0 & 0 & \psi_r & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \psi_r \end{bmatrix}$$

$$\underline{Y}_k = \begin{bmatrix} \psi_{k-1} & \psi_{k-2} & \cdot & \cdot & \psi_1 & \psi_0 \\ \psi_{k-2} & \psi_{k-3} & & & \psi_0 & 0 \\ \cdot & & \cdot & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \psi_0 & 0 & \cdot & \cdot & 0 & 0 \end{bmatrix}$$

where the ψ_s are the coefficients of $\Psi(z)$ given by (5.13). Jury then proves the following:

Theorem 5.3

When r is even, all the roots of the polynomial $\Psi(z)$ lie within the unit circle if and only if

(i) $\Psi(1) > 0$

(ii) $\Psi(-1) > 0$

(iii) $\det(\underline{X}_k \pm \underline{Y}_k) > 0 \qquad k = 1,3,\ldots r-1$.

We also have

Corollary

An alternative to (iii) is

(iii)' $\det(\underline{X}_k \pm \underline{Y}_k) > 0 \qquad k = 2,4,\ldots r-2$

and $\det(\underline{X}_{r-1} - \underline{Y}_{r-1}) > 0$.

Use of the corollary means only one determinant of order $r-1$ needs to be evaluated, while the theorem requires two such evaluations.

Theorem 5.4

When r is odd, all the roots of the polynomial $\Psi(z)$ lie within the unit circle if and only if

(i) $\Psi(1) > 0$

(ii) $\Psi(-1) < 0$

(iii) $\det(\underline{X}_k \pm \underline{Y}_k) > 0 \qquad k = 2,4,\ldots r-1$.

Again, there is an alternative to (iii), so that only one $(r-1)x(r-1)$ determinant needs to be evaluated

Corollary

An alternative to (iii) is

(iii)' $\det(\underline{X}_k \pm \underline{Y}_k) > 0 \qquad k = 1,3,\ldots,r-2$

and $\det(\underline{X}_{r-1} - \underline{Y}_{r-1}) > 0$.

Example 5.1

$$\alpha(z) = z^2+\alpha_1 z+\alpha_2$$

is required to have all roots inside the unit circle. Comparing this equation with (5.13) we see that $\psi_o = \alpha_2$, $\psi_1 = \alpha_1$, $\psi_2 = 1$. Using Theorem 5.3 and its corollary, the constraints are

$$1 + \alpha_1 + \alpha_2 > 0$$

$$1 - \alpha_1 + \alpha_2 > 0$$

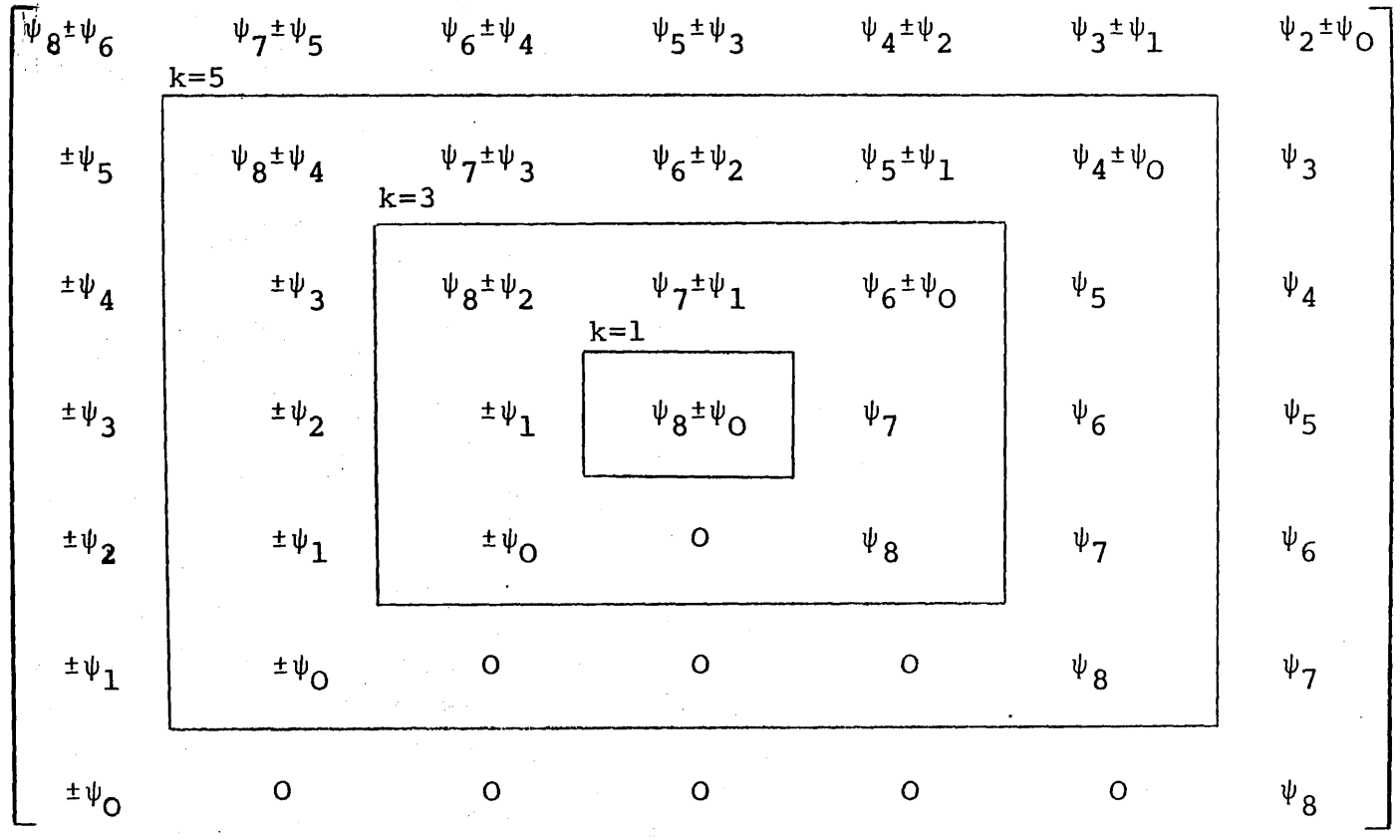
$$\det(\psi_2 - \psi_o) = 1-\alpha_2 > 0 .$$

These three inequalities specify completely the triangular stability region in (α_1,α_2) space.

In general, using the theorems without their corollaries, it is easily seen by writing out $(\underline{X}_{r-1} \pm \underline{Y}_{r-1})$ in full, that this matrix contains all the other $(\underline{X}_k \pm \underline{Y}_k)$ required. For example, when $r = 8$, $(\underline{X}_7 \pm \underline{Y}_7)$ is given by Table 5.1.

Jury's Table Method

We now consider the table method given by Jury for determining whether or not the roots of a polynomial are less than one in modulus. Given the polynomial



$\psi_8 \pm \psi_6$	$\psi_7 \pm \psi_5$	$\psi_6 \pm \psi_4$	$\psi_5 \pm \psi_3$	$\psi_4 \pm \psi_2$	$\psi_3 \pm \psi_1$	$\psi_2 \pm \psi_0$
$\pm\psi_5$	$\psi_8 \pm \psi_4$	$\psi_7 \pm \psi_3$	$\psi_6 \pm \psi_2$	$\psi_5 \pm \psi_1$	$\psi_4 \pm \psi_0$	ψ_3
$\pm\psi_4$	$\pm\psi_3$	$\psi_8 \pm \psi_2$	$\psi_7 \pm \psi_1$	$\psi_6 \pm \psi_0$	ψ_5	ψ_4
$\pm\psi_3$	$\pm\psi_2$	$\pm\psi_1$	$\psi_8 \pm \psi_0$	ψ_7	ψ_6	ψ_5
$\pm\psi_2$	$\pm\psi_1$	$\pm\psi_0$	0	ψ_8	ψ_7	ψ_6
$\pm\psi_1$	$\pm\psi_0$	0	0	0	ψ_8	ψ_7
$\pm\psi_0$	0	0	0	0	0	ψ_8

Table 5.1

$$\Psi(z) = \psi_o+\psi_1 z+\psi_2 z^2+\ldots+\psi_r z^r,$$

where $\psi_r > 0$, construct the following table,

Row	z^0	z^1	z^2	z^3		z^{r-k}		z^{r-2}	z^{r-1}	z^r
1	ψ_o	ψ_1	ψ_2	ψ_3	. . .	ψ_{r-k}	. . .	ψ_{r-2}	ψ_{r-1}	ψ_r
2	ψ_r	ψ_{r-1}	ψ_{r-2}	ψ_{r-3}	. . .	ψ_k	. . .	ψ_2	ψ_1	ψ_o
3	b_o	b_1	b_2	b_3				b_{r-2}	b_{r-1}	
4	b_{r-1}	b_{r-2}	b_{r-3}	b_{r-4}				b_1	b_o	
5	c_o	c_1	c_2	c_3				c_{r-2}		
6	c_{r-2}	c_{r-3}	c_{r-4}	c_{r-5}				c_o		
.	.	.	.	.						
.	.	.	.	.						
.	.	.	.	.						
.	.	.	.	.						
2r-5	r_o	r_1	r_2	r_3						
2r-4	r_3	r_2	r_1	r_o						
2r-3	s_o	s_1	s_2							
2r-2	s_2	s_1	s_o							
2r-1	t_o	t_1								

Row $2j$ is formed from row $2j-1$ by writing the elements in reverse order. The elements $b_k, c_k, \ldots, t_k$ are found from the appropriate 2x2 determinant from the preceeding two rows

i.e.
$$b_k = \begin{vmatrix} \psi_0 & \psi_{r-k} \\ \psi_r & \psi_k \end{vmatrix}, \quad c_k = \begin{vmatrix} b_0 & b_{r-1-k} \\ b_{r-1} & b_k \end{vmatrix},$$

$$t_o = \begin{vmatrix} s_0 & s_2 \\ s_2 & s_0 \end{vmatrix}, \quad t_1 = \begin{vmatrix} s_0 & s_1 \\ s_2 & s_1 \end{vmatrix}.$$

Theorem 5.5 (Jury)

All the roots of $\Psi(z)$ lie within the unit circle if and only if

$$\Psi(1) > 0$$
$$(-1)^r \Psi(-1) > 0$$
$$\left.\begin{aligned} b_o &< 0 \\ c_o &> 0 \\ d_o &> 0 \\ &\vdots \\ s_o &> 0 \\ t_o &> 0 \end{aligned}\right\} \quad (r-1) \text{ constraints.}$$

Example 5.2

$$\Psi(z) = 2+5z+7z^2+6z^3+3z^4$$
$$\Psi(1) = 2+5+7+6+3 > 0$$
$$(-1)^4 \Psi(-1) = 2-5+7-6+3 > 0.$$

Thus the first two conditions are satisfied. Setting up the table

Row	z^0	z^1	z^2	z^3	z^4
1	2	5	7	6	3
2	3	6	7	5	2
3	-5	-8	-7	-3	
4	-3	-7	-8	-5	
5	16	19	11		
6	11	19	16		
7	135	95			

From the above table

$$b_o = -5 \quad < 0$$

$$c_o = 16 \quad > 0$$

$$d_o = 135 \quad > 0 .$$

Thus all constraints are satisfied and the polynomial has all roots inside the unit circle

5.3 Observability

The concept of observability introduced by Kalman (1961) is an integral part of modern control theory. Many important theorems depend on this idea, and on its dual concept of controllability. Briefly a system or model can be defined as observable if every change of the state vector eventually affects the output (observation). To fix ideas, consider first the purely deterministic system

$$\begin{aligned} \underline{Y}_t &= \underline{F}_t \underline{\theta}_t \\ \underline{\theta}_t &= \underline{G}_t \underline{\theta}_{t-1} \end{aligned} \tag{5.14}$$

which is the model (4.1) with the random terms missing. As before, $\underline{y}_t$ is mxl, $\underline{\theta}_t$ is nxl. The model is said to be completely observable if for any t_o there is a finite $t_j > t_o$ such that every state $\underline{\theta}_{t_o}$ can be expressed as a function of the observations $\underline{y}_{t_o} \cdots \underline{y}_{t_j}$. Because of the deterministic nature of the model, this is equivalent to expressing $\underline{\theta}_k$ at some time $k > 0$ in terms of the observations $\underline{y}_o, \underline{y}_1, \cdots \underline{y}_k$.

Let

$$\begin{aligned} \underline{Y} &= (\underline{y}_o^T, \underline{y}_1^T, \ldots, \underline{y}_k^T)^T \\ &= \left[(\underline{F}_o\underline{\theta}_o)^T, (\underline{F}_1\underline{\theta}_1)^T, \ldots, (\underline{F}_k\underline{\theta}_k)^T\right]^T \\ &= \underline{M}_k\underline{\theta}_o \end{aligned} \tag{5.15}$$

where

$$\underline{M}_k = \begin{bmatrix} \underline{F}_o \\ \underline{F}_1\underline{G}_1 \\ \underline{F}_2\underline{G}_2\underline{G}_1 \\ \\ \underline{F}_k \prod_{j=1}^{k} \underline{G}_j \end{bmatrix} . \tag{5.16}$$

We assume $m(k+1) \geq n$, for we cannot expect to learn much about an n dimensional parameter from less than n observations.

From the theory of the general linear model, we know that if the $m(k+1)xn$ matrix $\underline{M}_k$ is of full rank, then

$$\underline{\theta}_o = (\underline{M}_k^T\underline{M}_k)^{-1}\underline{M}_k^T\underline{Y} . \tag{5.17}$$

The inverse matrix exists if and only if $\text{rank}(\underline{M}_k)$ is n. Thus it follows that the model is observable if and only if $\text{rank}(\underline{M}_k)$ is n. When $\underline{F}_t$ and $\underline{G}_t$ are both time independent, this observability criterion reduces to

$$\left[\underline{F}^T \ (\underline{FG})^T \ldots (\underline{FG}^k)^T\right] \quad \text{must have rank} \quad n \,. \tag{5.18}$$

The characteristic equation of the matrix $\underline{G}$ has degree n so, by the Cayley-Hamilton theorem that every square matrix satisfies its own characteristic equation, $\underline{FG}^j$ can be expressed in terms of $\underline{F}, \underline{FG}, \ldots, \underline{FG}^{n-1}$ for every $j \geq n$. Thus the observability criterion (5.18) reduces to

$$\text{rank}\left[\underline{F}^T (\underline{FG})^T \ldots (\underline{FG}^{n-1})^T\right] = n \,. \tag{5.19}$$

When $m > 1$, so that k could be less than $n-1$, it seems possible that (5.19) could be satisfied while (5.18) is not. However, the criterion for observability given by Jacobs (1974) is that the set of matrices $\{\underline{F}, \underline{FG}, \ldots, \underline{FG}^{n-1}\}$ is a linearly independent set, which is the same as criterion (5.19). The matrix in (5.19) is often called the observability matrix.

In the time varying case, it is clearly not possible to simplify the criterion in this way. Returning to equation (5.17), an alternative criterion for observability is that the matrix $\underline{M}_k^T \underline{M}_k$ should be non-singular, or, assuming for simplicity that $\underline{G}$ is constant, the quantity

$$\sum_{i=0}^{k} (\underline{G}^i)^T \underline{F}_i^T \underline{F}_i \underline{G}^i \tag{5.20}$$

is positive definite.

Now consider the model with observation noise, but still without system error. For simplicity, we also assume $\underline{G}$ is constant. The model is given by

$$\begin{aligned} \underline{Y}_t &= \underline{F}_t \underline{\theta}_t + \underline{\varepsilon}_t \\ \underline{\theta}_t &= \underline{G}\underline{\theta}_{t-1} \end{aligned} \tag{5.21}$$

where $E\left[\underline{\varepsilon}_t\right] = \underline{0} \quad E\left[\underline{\varepsilon}_t \underline{\varepsilon}_{t+k}\right] = \delta_{o,k} \underline{V}_t$.

It is no longer possible to determine $\underline{\theta}_i$ exactly from the observations, but we can of course, estimate $\underline{\theta}_i$ using the probability density function $p(\underline{\theta}_i | \underline{Y}_0, \underline{Y}_1, \ldots \underline{Y}_i)$.

Again, if at time t we can estimate the state at one time, say $\underline{\theta}_0$, then we can estimate $\underline{\theta}_i$ for all $i \leq t$. If $\underline{Y} = (\underline{Y}_0^T, \underline{Y}_1^T, \ldots, \underline{Y}_k^T)^T$ as before, then

$$\underline{Y} = \underline{M}_k \underline{\theta}_0 + \underline{e}_k \tag{5.22}$$

where

$$\underline{M}_k = \begin{bmatrix} \underline{F}_0 \\ \underline{F}_1\underline{G} \\ \underline{F}_2\underline{G}^2 \\ \vdots \\ \underline{F}_k\underline{G}^k \end{bmatrix} \quad \text{and} \quad \underline{e}_k \begin{bmatrix} \underline{\varepsilon}_0 \\ \underline{\varepsilon}_1 \\ \vdots \\ \underline{\varepsilon}_k \end{bmatrix} .$$

Since $E\left[\underline{\varepsilon}_k \underline{\varepsilon}_j^T\right] = \underline{0}$ for $k \neq j$,

$$E\left[\underline{e}_k \underline{e}_k^T\right] = \begin{bmatrix} \underline{V}_0 & & & & \\ & \underline{V}_1 & & & \\ & & \underline{V}_2 & & \\ & & & \ddots & \\ & & & & \underline{V}_k \end{bmatrix} = \underline{V}_k .$$

From the theory of the generalised linear model (see, for example, Goldberger, Chapter 5), the minimum variance linear unbiased estimator of $\underline{\theta}_0$ is given by

$$\hat{\underline{\theta}}_0 = (\underline{M}_k^T \underline{V}_k^{-1} \underline{M}_k)^{-1} \underline{M}_k^T \underline{V}_k^{-1} \underline{Y} . \tag{5.23}$$

If this estimator is to exist, then $\underline{M}_k^T \underline{V}_k^{-1} \underline{M}_k$ must be positive definite, or

$$\sum_{i=0}^{k} (\underline{G}^i)^T \underline{F}_i^T \underline{V}_i^{-1} \underline{F}_i \underline{G}^i \tag{5.24}$$

is positive definite. In particular, when the observations are scalar and the observation noise is the same for all i, so that $v_i^{-1} = 1/v$ for all i, then the condition (5.24) reduces to (5.20), and if $\underline{F}_i$ is also constant, it further reduces to (5.19). The matrix (5.24) is also sometimes called the observability matrix, or the information matrix.

If we now add plant noise, i.e. w_t to the model (5.21), we have returned to the DLM (4.17). In this case, knowledge of $\underline{\theta}_i$ at any time point i does not determine any other $\underline{\theta}_j$ exactly, because of the plant noise. Observability of this model is considered in several texts, for example, Aoki (1967, p.209-221). From the general definition of observability, as given at the beginning of this section, the model is found to be observable if two subsystems are observable in the sense described above. Thus, even in this most general case, many of the important theorems rely only on observability in the sense of the deterministic model, i.e. it is merely required that $\text{rank}(\underline{M}_k)$ is n, where $\underline{M}_k$ is defined in equation (5.16). In most of what follows, we shall be concerned only with this simpler criterion. Until the end of Chapter 9, observability of the model will mean observability of the appropriate deterministic model. Furthermore, we shall prefer to assume $\underline{F}$ and $\underline{G}$ constant. In Chapter 10, however, we shall return to the model (5.21) and time varying $\underline{F}$ and $\underline{G}$.

5.4 Some consequences of observability

It is helpful for what follows to look at the asymptotic properties of the model (4.1) and the estimation method defined by the Kalman filter algorithm given in equations (4.4) to (4.8). In many cases, the filter settles to an equilibrium state, where the Kalman gain matrix $\underline{A}_t$ and the covariance matrix $\underline{C}_t$ are constant. The comparison of the dynamic Kalman filter approach with the static ARIMA approach is only applicable when the estimation scheme converges to one which is independent of time. For this to be possible, $\underline{F}$, $\underline{G}$, $\underline{H}$, $\underline{V}$ and $\underline{W}$ must be independent of time, and the time interval between observations must be constant. Given this situation, sufficient (but not necessary) conditions for the covariance matrix $\underline{C}_t$ (and hence the Kalman gain matrix) to converge to a steady state value are given by Kalman (1963a). We shall express this result in a form due essentially to Kushner (1971, §9.5).

Theorem 5.6

Let the model be given by (4.1), with $\underline{H}_t = \underline{I}$ and $\underline{F}$, $\underline{G}$, $\underline{V}$ and $\underline{W}$ independent of t, and let the system covariance matrix $\underline{W}$ be written $\underline{M}\underline{M}^T$. If the model is observable, and if the DLM

$$\begin{aligned} \underline{z}_t &= \underline{M}^T \underline{\phi}_t + \underline{\varepsilon}_t \\ \underline{\phi}_t &= \underline{G}^T \underline{\phi}_{t-1} \end{aligned} \qquad (5.25)$$

is observable, then the covariance matrix

$$\underline{C}_t = E\left[(\underline{\theta}_t - \hat{\underline{\theta}}_t)(\underline{\theta}_t - \hat{\underline{\theta}}_t)^T\right]$$

converges to a constant matrix

$$\underline{C} = \underline{P} - \underline{PF}^T(\underline{FPF}^T + \underline{V}^{-1})\underline{FP}$$

where

$$\underline{P} = \underline{GCG}^T + \underline{W}$$

independently of the initialisation. This $\underline{C}$ is unique and positive definite.

Corollary

If $\underline{W}$ is positive definite and the model is observable, then $\underline{C}_t$ converges to a constant matrix as above.

For if $\underline{W}$ is positive definite, then $\underline{M}$ has rank n, thus the model (5.25) is observable, and the result follows from the theorem.

It should be noted that if a DLM does not satisfy the conditions of the theorem, it could still have an equilibrium solution. There do not appear to be any known necessary and sufficient conditions for the existence of a steady state solution. However, since the steady state is essential to the major part of this work, we shall require all our models to be observable. Since we do not wish to investigate the system error covariance matrix, we shall also assume that this matrix is positive definite. Then we can assume the equilibrium model and apply z-transforms, as described in Section 5.1.

CHAPTER 6

THE CONSTANT FORECAST MODEL

6.1 Discussion

The constant forecast model is defined by the fact that the predictor at time t is constant for all lags, i.e.

$$y_t(k) = y_t(1) \quad \text{for all} \quad k \geq 1. \tag{6.1}$$

Models satisfying (6.1) have the appearance of being 'trend free' and hence have a special place in prediction theory. They are among the most widely used models in forecasting, for a 'trend free' or steady model often appears to be an adequate description of time series data in the short term. Many representations exist, notably those of Holt (1957), Muth (1960), Brown (1962), Box and Jenkins (1970) and Harrison and Stevens (1976). Several authors have looked in detail at these models; recently these have included Smith (1979), Godolphin and Stone (1980) and Key and Godolphin (1981).

We shall investigate the properties of the DLMs proposed by Harrison and Stevens (1976). In Section 4.3, the predictor $y_t(k)$ for a DLM was defined as the expectation of y_{t+k} based on the observations up to time t. Thus in DLM notation, we have

$$y_t(k) = \underline{F}\underline{G}^k\hat{\underline{\theta}}_t \tag{6.2}$$

so that from (6.1)

$$\underline{F}\underline{G}^k\hat{\underline{\theta}}_t = \underline{F}\underline{G}\hat{\underline{\theta}}_t \quad \text{for} \quad k \geq 1. \tag{6.3}$$

Since this should hold for all values of $\hat{\underline{\theta}}_t$, we require

$$\underline{FG}^k = \underline{FG} \quad \text{for} \quad k \geq 1$$

and in particular

$$\underline{FG}^2 = \underline{FG}. \tag{6.4}$$

Any DLM satisfying (6.4) is a constant forecast model, i.e. (6.1) will be satisfied for all $k \geq 1$.

6.2 Examples of Constant Forecast Models

Example 6.1

The natural model to use for this situation appears to be the purely scalar model, i.e. $n = 1$ and

$$\begin{aligned} y_t &= u_t + \varepsilon_t \\ u_t &= u_{t-1} + w_t \end{aligned} \tag{6.5}$$

where ε_t, w_t are uncorrelated random variables with means zero. This example and its predictor are discussed in some detail by Harrison (1967), and it is the steady model used by Harrison and Stevens (1976).

Using equations (4.4) - (4.8), the Kalman updating equation for the model (6.5) is given by

$$\hat{u}_t = \hat{u}_{t-1} + A_t(y_t - \hat{u}_{t-1}) \tag{6.6}$$

where

$$A_t = (C_{t-1} + \text{var}(w_t))/(C_{t-1} + \text{var}(w_t) + \text{var}(\varepsilon_t)) \tag{6.7}$$

and

$$C_t = \text{var}(u_t - \hat{u}_t) = A_t \text{var}(\varepsilon_t). \tag{6.8}$$

Clearly

$$0 < A_t < 1 \quad \text{for all} \quad t. \tag{6.9}$$

Since $F = G = 1$, the predictor of y_{t+k} at time t is $\hat{u}_t$, with prediction error variance increasing steadily with k. It is assumed that $\text{var}(w_t)$, $\text{var}(\varepsilon_t)$ are independent of time, and that the observations are made at regular intervals. Since this model is observable and "W" is positive definite, then by Theorem 5.6, there is an equilibrium state where $C_t = C$, and hence $A_t = A$ for all large enough t. For this example, this can be verified directly as follows:

From equation (6.7) and (6.8)

$$A_t = 1 - \text{var}(\varepsilon_t)/(A_{t-1}\text{var}(\varepsilon_t) + \text{var}(w_t) + \text{var}(\varepsilon_t))$$

so that

$$(1-A_t)^{-1} = A_{t-1} + 1 + \text{var}(w_t)/\text{var}(\varepsilon_t).$$

Hence

$$(1-A_t)^{-1} - A_{t-1} = 1 + \text{var}(w_t)/\text{var}(\varepsilon_t)$$

$$(1-A_{t-1})^{-1} - A_{t-2} = 1 + \text{var}(w_t)/\text{var}(\varepsilon_t).$$

Subtracting, we find that

$$(1-A_t)^{-1} - (1-A_{t-1})^{-1} = A_{t-1} - A_{t-2}$$

so that

$$\frac{A_{t-1} - A_{t-2}}{A_t - A_{t-1}} = \frac{1}{(1-A_t)(1-A_{t-1})} . \qquad (6.10)$$

Since $0 < (1-A_t) < 1$ for all t from (6.9), the right hand side of (6.10) is positive, so that $A_t - A_{t-1}$ has the same sign as $A_{t-1} - A_{t-2}$. Thus the sequence A_t is monotonic, and is bounded above and below, therefore it tends to a limit, A. The limit must satisfy

$$A = 1 - \text{var}(\varepsilon_t)/(A\text{var}(\varepsilon_t) + \text{var}(w_t) + \text{var}(\varepsilon_t))$$

or

$$A^2 \mathrm{var}(\varepsilon_t) + A\mathrm{var}(w_t) - \mathrm{var}(w_t) = 0$$

which has exactly one positive solution

$$A = [-\mathrm{var}(w_t) + \{\mathrm{var}(w_t)(\mathrm{var}(w_t)+4\mathrm{var}(\varepsilon_t))\}^{\frac{1}{2}}]/2\mathrm{var}(\varepsilon_t).$$

Thus for large t, (6.6) becomes

$$\hat{u}_t = \hat{u}_{t-1} + A(y_t - \hat{u}_{t-1})$$

or

$$\hat{u}_t - (1-A)\hat{u}_{t-1} = Ay_t.$$

This is the familiar expression for the EWMA, which for $|1-A| < 1$, has solution

$$\hat{u}_t = A \sum_{j=0}^{\infty} (1-A)^j y_{t-j} . \qquad (6.11)$$

From (6.9) it follows that $\hat{u}_t$ assumes only a subset of its possible values, since it restricted by $0 < A < 1$, whereas the series (6.11) is valid for $0 < A < 2$. Thus any constant forecast model should allow A to take all values in the range $0 < A < 2$. Godolphin (1976) pointed out that this can be accomplished for the Example 6.1 by the introduction of covariances, but this means that the model is no longer strictly a DLM. For example, set $\mathrm{cov}(w_t,\varepsilon_t) = x \neq 0$. Then the Kalman updating equations (4.4) - (4.8) cannot be applied as they stand, since they rely on the independence of system and observation noise. We shall derive the updating formulae for this dependent case. $\hat{u}_t$ will now be given by

$$\hat{u}_t = \hat{u}_{t-1} + K_t(y_t - \hat{u}_{t-1})$$

where K_t is chosen to minimise

$$C_t = E\left[(u_t - \hat{u}_t)^2\right]$$

$$= E\left[(u_{t-1}+w_t-\hat{u}_{t-1}-K_t(u_t+\varepsilon_t-\hat{u}_{t-1}))^2\right]$$

$$= (1-K_t)^2(C_{t-1}+\text{var}(w_t)) - 2K_t(1-K_t)\text{cov}(w_t,\varepsilon_t)$$

$$+ K_t^2\text{var}(\varepsilon_t)). \quad (6.12)$$

Differentiating this equation for C_t with respect to K_t

$$\frac{\partial C_t}{\partial K_t} = -2(1-K_t)(C_{t-1}+\text{var}(w_t)) - 2(1-2K_t)\text{cov}(w_t,\varepsilon_t)$$

$$+ 2K_t\text{var}(\varepsilon_t).$$

For a minimum, $\frac{\partial C_t}{\partial K_t} = 0$, thus

$$-C_{t-1}-\text{var}(w_t)-\text{cov}(w_t,\varepsilon_t)$$

$$+ K_t[C_{t-1}+\text{var}(w_t)+2\text{cov}(w_t,\varepsilon_t) + \text{var}(\varepsilon_t)] = 0$$

or

$$K_t = \frac{C_{t-1}+\text{var}(w_t)+\text{cov}(w_t,\varepsilon_t)}{C_{t-1}+\text{var}(w_t+\varepsilon_t)} \quad . \quad (6.13)$$

Substituting this value of K_t into (6.12) yields

$$C_t = \frac{\text{var}(\varepsilon_t)(C_{t-1}+\text{var}(w_t))-(\text{cov}(w_t,\varepsilon_t))^2}{C_{t-1}+\text{var}(w_t+\varepsilon_t)} \quad . \quad (6.14)$$

It appears that K_t can be positive or negative, depending on the value of $\text{cov}(w_t,\varepsilon_t)$. It remains to be shown that, at least in the steady state situation, it can take any value in the range $0 < K_t < 2$. C_t is positive for all values of $\text{cov}(w_t,\varepsilon_t)$.

Assuming the existence of a steady state solution, then for large enough t

$$C = \frac{\text{var}(\varepsilon_t)(C+\text{var}(w_t)) - (\text{cov}(w_t,\varepsilon_t))^2}{C+\text{var}(w_t+\varepsilon_t)}$$

or

$$C^2+C(\text{var}(w_t)+2\text{cov}(w_t,\varepsilon_t)) - \{\text{var}(w_t)\text{var}(\varepsilon_t)-(\text{cov}(w_t,\varepsilon_t))^2\} = 0.$$

This has solutions

$$C = -\tfrac{1}{2}(\text{var}(w_t) + 2\text{cov}(w_t,\varepsilon_t))$$

$$\pm\tfrac{1}{2}\{\left[\text{var}(w_t)+2\text{cov}(w_t,\varepsilon_t)\right]^2 + 4\left[\text{var}(w_t)\text{var}(\varepsilon_t)-(\text{cov}(w_t,\varepsilon_t))^2\right]\}^{\frac{1}{2}}.$$

These two solutions are real and of opposite sign, hence the admissible positive solution is

$$C = -\tfrac{1}{2}(\text{var}(w_t)+2\text{cov}(w_t,\varepsilon_t))+\tfrac{1}{2}\{\text{var}(w_t)\text{var}(w_t+2\varepsilon_t)\}^{\frac{1}{2}}.$$

Substituting this value into the expression for $1-K$,

$$1-K = (\text{cov}(w_t,\varepsilon_t)+\text{var}(\varepsilon_t))/(C+\text{var}(w_t+\varepsilon_t))$$

$$= \frac{2(\text{cov}(w_t,\varepsilon_t) + \text{var}(\varepsilon_t))}{\text{var}(w_t+\varepsilon_t)+\text{var}(\varepsilon_t)+\{\text{var}(w_t)\text{var}(w_t+2\varepsilon_t)\}^{\frac{1}{2}}}.$$

As $\text{var}(w_t) \to 0$, $1 - K \to 1$.

As $\text{var}(w_t+2\varepsilon_t) \to 0$, $1 - K \to -1$.

Thus the predictor of a model of this form is equivalent in the steady state to that of any ARIMA (0,1,1) model, since $0 < K < 2$. However, as previously stated, the Kalman filter equations do not apply strictly when the observation and system noise are dependent, and the above method of deriving the relevant equations is somewhat more difficult for models of larger dimension. We therefore write the model in another form, so that the standard Kalman filter equations can be used.

Example 6.2

Consider

$$\begin{aligned} y_t &= u_t + \phi_t \\ u_t &= u_{t-1} + w_t \\ \phi_t &= \varepsilon_t . \end{aligned} \tag{6.15}$$

It is easily seen that this model is effectively the same as (6.5), despite the introduction of the extra "state variable" ϕ_t. It has dimension two, with

$$\underline{F} = \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad \underline{G} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \tag{6.16}$$

and

$$\underline{W}_t = \begin{bmatrix} \mathrm{var}(w_t) & \mathrm{cov}(w_t, \varepsilon_t) \\ \mathrm{cov}(w_t, \varepsilon_t) & \mathrm{var}(\varepsilon_t) \end{bmatrix}$$

Notice the absence of observation noise. Clearly, $\underline{FG}^k = \underline{FG} = \begin{bmatrix} 1 & 0 \end{bmatrix}$ for all $k \geq 1$, so that (6.15) is a constant forecast model.

Applying the Kalman filter equations (4.4) - (4.8), we find that

$$\begin{bmatrix} \hat{u}_t \\ \hat{\phi}_t \end{bmatrix} = \begin{bmatrix} \hat{u}_{t-1} \\ 0 \end{bmatrix} + \underline{A}_t (y_t - \hat{u}_{t-1}) \tag{6.17}$$

with $\underline{A}_t = \begin{bmatrix} A_{1t}, & A_{2t} \end{bmatrix}^T$ given by

$$A_{1t} = \frac{c_{t-1} + \mathrm{var}(w_t) + \mathrm{cov}(w_t, \varepsilon_t)}{c_{t-1} + \mathrm{var}(w_t + \varepsilon_t)} \quad \text{and} \quad A_{2t} = \frac{\mathrm{cov}(w_t, \varepsilon_t) + \mathrm{var}(\varepsilon_t)}{c_{t-1} + \mathrm{var}(w_t + \varepsilon_t)}$$

and

$$\underline{C}_t = c_t \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

where

$$c_t = \frac{\mathrm{var}(\varepsilon_t)(c_{t-1} + \mathrm{var}(w_t)) - (\mathrm{cov}(w_t, \varepsilon_t))^2}{c_{t-1} + \mathrm{var}(w_t + \varepsilon_t)} . \tag{6.18}$$

Since we are primarily interested in the predictor of y_{t+k}, $\underline{FG}^k \hat{\underline{\theta}}_t$, we premultiply equation (6.17) by $\underline{FG}^k = \begin{bmatrix} 1 & 0 \end{bmatrix}$.

Thus

$$\hat{u}_t = \hat{u}_{t-1} + A_{1t}(y_t - \hat{u}_{t-1}) \qquad (6.19)$$

with A_{1t} and c_t as for the model (6.5). As expected, the predictor of this 2 x 2 DLM is the same as the predictor of the scalar model with dependent system and observation noise. This seems to be an improvement, because we can now use the standard Kalman filter equations, but if the 'smoothing constant' is to be able to take its full range of values, then we need to specify covariances between the elements of the system error matrix. However, the Kalman filter is sensitive to the matrices V_t and $\underline{W}_t$, so it is important that they are accurately determined. It seems likely that even experienced forecasters who can usually specify means and variances, and whether there is any correlation between the state variables, may well be unable to evaluate the covariances accurately. Hence it is advisable to find a model for which it is possible to specify a diagonal covariance matrix and still enable the Kalman gain matrix for the predictor, FGA, to take all values in the range $0 < \underline{FGA} < 2$.

We investigate several examples in an attempt to find such a model.

Example 6.3

Consider

$$y_t = \begin{bmatrix} a & 0 \end{bmatrix} \underline{\theta}_t + v_t$$

$$\underline{\theta}_t = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \underline{\theta}_{t-1} + \begin{bmatrix} w_{1t} \\ w_{2t} \end{bmatrix} \qquad (6.20)$$

where

$$\underline{W} = \begin{bmatrix} w_1 & 0 \\ 0 & w_2 \end{bmatrix}$$

is diagonal and independent of time, $v = \mathrm{var}(v_t)$ and

a is a fixed but arbitrary positive number.

Applying the Kalman updating procedure,

$$\underline{\hat{\theta}}_t = \underline{G}\underline{\hat{\theta}}_{t-1} + \underline{A}_t(y_t - a\hat{\theta}_{1t-1}) \qquad (6.21)$$

where

$\underline{A}_t = \begin{bmatrix} A_{1t} & A_{2t} \end{bmatrix}^T$ is given by

$$A_{1t} = a(c_{t-1} + w_1)/(a^2(c_{t-1}+w_1)+v)$$

$$A_{2t} = ac_{t-1}/(a^2(c_{t-1}+w_1)+v)$$

and

$$c_t = \mathrm{var}(\theta_{1t} - \hat{\theta}_{1t}).$$

We now turn our attention to the predictor of the model (6.20), by premultiplying equation (6.21) by $\underline{FG} = \begin{bmatrix} a & 0 \end{bmatrix}$

$$a\hat{\theta}_{1t} = a\hat{\theta}_{1t-1} + aA_{1t}(y_t - a\hat{\theta}_{1t-1})$$

so that the 'smoothing constant' $\mu_t = \underline{FGA}_t = aA_{1t}$ is

$$\mu_t = \frac{a^2(c_{t-1}+w_1)}{a^2(c_{t-1}+w_1)+v} .$$

Thus μ_t lies between 0 and 1 whatever the value of a. Indeed, as $v \to 0$, $\mu_t \to 1$, however small a may be. We conclude that this model is no more general for forecasting purposes than the previous one.

Example 6.4

Now consider a DLM with more complicated structure

$$\begin{aligned} y_t &= \tfrac{1}{2}\begin{bmatrix} \frac{1+a}{a} & a-1 \end{bmatrix}\underline{\theta}_t + v_t \\ \underline{\theta}_t &= \tfrac{1}{2}\begin{bmatrix} 1 & a \\ 1/a & 1 \end{bmatrix}\underline{\theta}_{t-1} + \underline{w}_t \end{aligned} \qquad (6.22)$$

where $\mathrm{var}(v_t) = v$ and

$$\underline{W} = \begin{bmatrix} w_1 & 0 \\ 0 & w_2 \end{bmatrix}$$

are time independent and $a > 0$ as before.

Applying the Kalman updating procedure

$$\hat{\underline{\theta}}_t = \underline{G}\hat{\underline{\theta}}_{t-1} + \underline{A}_t(y_t - \tfrac{1}{2}(\theta_{1t-1} + a\theta_{2t-1})), \tag{6.23}$$

where $\underline{A}_t = \left[A_{1t}\ A_{2t}\right]^T$ is given by

$$A_{1t} = \left(a^2 p_{t-1} + 2\frac{(1+a)}{a} w_1\right)/D_t$$

$$A_{2t} = \left(a p_{t-1} + 2(a-1)w_2\right)/D_t.$$

Here

$$D_t = a^2 p_{t-1} + \frac{(1+a)^2}{a^2} w_1 + (a-1)^2 w_2 + v$$

and

$$p_t = \mathrm{var}((\theta_{1t} - \hat{\theta}_{1t})/a + (\theta_{2t} - \hat{\theta}_{2t})).$$

Turning our attention to the predictor, by premultiplying equation (6.23) by $\underline{FG} = \tfrac{1}{2}\left[1\ a\right]$, and writing

$$m_t = \tfrac{1}{2}(\theta_{1t} + a\theta_{2t}),\ \mu_t = \tfrac{1}{2}(A_{1t} + aA_{2t})$$

we find that

$$m_t = m_{t-1} + \mu_t(y_t - m_{t-1})$$

where μ_t is given by

$$\mu_t = \left(a^2 p_{t-1} + \frac{(1+a)}{a} w_1 + a(a-1)w_2\right)/D_t$$

or

$$1 - \mu_t = \left(\frac{1+a}{a^2} w_1 + (1-a)w_2\right)/D_t. \tag{6.24}$$

It is clear that for $0 < a \leq 1$, $1 - \mu_t$ is positive, while for $a > 1$, $1 - \mu_t$ can be negative if

$$w_2 > \frac{1+a}{a^2(a-1)} w_1 .$$

For example, when $a = 1$,

$$1-\mu_t = 2w_1/(p_{t-1} + 4w_1 + v)$$

so that

$$0 < 1 - \mu_t < \tfrac{1}{2}.$$

When $a = \frac{1}{2}$,

$$1-\mu_t = (6w_1+\tfrac{1}{2}w_2)/(\tfrac{1}{4}p_{t-1}+9w_1+\tfrac{1}{4}w_2+v)$$

so that

$$0 < 1 - \mu_t < 2.$$

Thus when $a = \frac{1}{2}$, the model (6.22) does not have its 'smoothing constant' within the bounds specified by the ARIMA (0,1,1) model for all t. This bound, however, was only specified in the steady state, and it will be shown in Section 6.3 that because of the value of p, the equilibrium value of μ lies in the range $0 < \mu < 2$ as required. When $a = 2$,

$$1-\mu_t = (\tfrac{3}{4}w_1-w_2)/(4p_{t-1}+9w_1/4+w_2+v)$$

which takes values between -1 and $1/3$.

Thus for the model (6.22), the quantity a is crucial to the range of forecasts that can be achieved, for as the value of a changes, the range of values taken by μ_t changes accordingly.

Example 6.5

A similar but somewhat simpler DLM is given by

$$y_t = \begin{bmatrix} 1+a & 1 \end{bmatrix} \underline{\theta}_t + v_t \qquad (6.25)$$

$$\underline{\theta}_t = \begin{bmatrix} 1 & 0 \\ -a & 0 \end{bmatrix} \underline{\theta}_{t-1} + \underline{w}_t$$

where $\text{var}(v_t) = v$ and

$$\underline{W} = \begin{bmatrix} w_1 & 0 \\ 0 & w_2 \end{bmatrix}$$

are time independent, and a is any real number. When $a = 0$, this reduces to Example 6.2.

Applying the Kalman updating equations

$$\hat{\underline{\theta}}_t = \underline{G}\hat{\underline{\theta}}_{t-1} + \underline{A}_t(y_t - \hat{\theta}_{1t-1}) \tag{6.26}$$

where $\underline{A}_t = \left[A_{1t}\ A_{2t}\right]^T$ is given by

$$A_{1t} = (c_{t-1}+(1+a)w_1)/(c_{t-1}+(1+a)^2w_1+w_2+v)$$

$$A_{2t} = (-ac_{t-1} + w_2)/(c_{t-1}+(1+a)^2w_1+w_2+v)$$

and

$$c_t = \text{var}(\theta_{1t} - \hat{\theta}_{1t}).$$

The corresponding predictor equation, found by premultiplying equation (6.26) by $\underline{F}\underline{G}^k = \left[1\ 0\right]$ is

$$\hat{\theta}_{1t} = \hat{\theta}_{1t-1} + A_{1t}(y_t - \hat{\theta}_{1t-1})$$

$$1-A_{1t} = \frac{a(1+a)w_1+w_2+v}{c_{t-1}+(1+a)^2w_1+w_2+v}. \tag{6.27}$$

The values taken by $\mu_t = A_{1t}$ clearly depend on a, as well as the values of w_1, w_2, v and c_t. When $a > 0$ or $a \leq -1$, then $1 - A_{1t}$ is positive for all t, but if $-1 < a < 0$, then it is possible that $1 - A_{1t}$ is negative. Consider some particular values.

(i) $a = 1$

$1-A_{1t} = (2w_1+w_2+v)/(c_{t-1}+4w_1+w_2+v)$

which lies between 0 and 1 for all t.

(ii) $a = -2$

$1-A_{1t} = (2w_1+w_2+v)/(c_{t-1}+w_1+w_2+v)$

which lies between 0 and 2 for all t.

(iii) $a = -\frac{1}{4}$

$1-A_{1t} = (-3w_1/16+w_2+v)/(c_{t-1}+9w_1/16+w_2+v)$

In this case $-1/3 < 1 - A_{1t} < 1$ for all t.

(iv) $a = -2/3$

$$1-A_{1t} = (-2w_1/9+w_2+v)/(c_{t-1}+w_1/9+w_2+v).$$

Here $-2 < 1 - A_{1t} < 1$ for all t.

On closer consideration of equation (6.27) we find that if $(1-A_{1t})$ can take positive values only ($a \leq -1$, or $a \geq 0$), then the lower bound on $1-A_{1t}$ is 0 and the upper bound is $\max(1, a/(1+a))$.

For $a > 0$, the maximum is one, as in (i), but for $a < -1$, the upper bound is strictly greater than one, as in (ii).

When $1-A_{1t}$ can take negative values, then the upper bound is one, and the lower bound is $a/(1+a)$ which is negative for $-1 < a < 0$. It is greater than -1 only for $-\frac{1}{2} \leq a < 0$, hence (iv) can fall outside the region $(-1,1)$, whilst (iii) lies within it.

Although the region $(-1,1)$ has only been derived for the steady state, it seems desirable that the parameter should lie within the region for all t, which means $a \geq -\frac{1}{2}$. It seems preferable that it should cover the whole region, and this requires $a = -\frac{1}{2}$.

The steady state solution is not difficult to derive for the model (6.25); after some straightforward algebra, we obtain

$$c = -\tfrac{1}{2}w_1(1+2a) + \tfrac{1}{2}\{w_1^2(1+2a)^2 + 4w_1(w_2+v)\}^{\frac{1}{2}}$$

hence

$$1-A_1 = \frac{2\left[a(1+a)w_1+w_2+v\right]}{(1+2a+2a^2)w_1+2(w_2+v)+\{w_1^2(1+2a)^2+4w_1(w_2+v)\}^{\frac{1}{2}}}$$

which lies between -1 and 1. It is clear that as $w_1 \to 0$, $1-A_1$ approaches 1. It is not so clear that the lower

bound of -1 can be attained. To find the minimum of $1-A_1$ for different values of a, we differentiate the expression for $1-A_1$ with respect to a, and set the result equal to zero. The numerator of the derivative is

$$2(1+2a)w_1\left[(1-2w_1a(1+a)/r)w_1-2(w_2+v)w_1/r+r\right]$$

where

$$r = \{w_1^2(1+2a)^2+4w_1(w_2+v)\}^{\frac{1}{2}}.$$

This is zero when $a = -\frac{1}{2}$, and it can be shown that this is a minimum.

Hence if $a = -\frac{1}{2}$, the model (6.25) can cover the full range of values $0 < A_1 < 2$ in both the time dependent and the equilibrium case with $\underline{W}$ diagonal. We consider this model in even more detail. Writing out (6.25) in full

$$\begin{aligned} y_t &= \tfrac{1}{2}\theta_{1t} + \theta_{2t} + v_t \\ \theta_{1t} &= \theta_{1t-1} + w_{1t} \\ \theta_{2t} &= \tfrac{1}{2}\theta_{1t-1} + w_{2t}. \end{aligned}$$

Let $\theta_{3t} = \frac{1}{2}\theta_{1t}$. Then we can write the model as

$$\begin{aligned} y_t &= \theta_{3t} + \theta_{2t} + v_t \\ \theta_{3t} &= \theta_{3t-1} + w_{3t} \\ \theta_{2t} &= \theta_{3t-1} + w_{2t}. \end{aligned} \tag{6.28}$$

Thus we have

Example 6.6

$$y_t = \begin{bmatrix} 1 & 1 \end{bmatrix} \underline{\theta}_t + v_t$$

$$\underline{\theta}_t = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \underline{\theta}_{t-1} + \begin{bmatrix} \gamma_{1t} \\ \gamma_{2t} \end{bmatrix} .$$

Let

$$E\left[\begin{bmatrix} \gamma_{1t} \\ \gamma_{2t} \end{bmatrix} \begin{bmatrix} \gamma_{1t} & \gamma_{2t} \end{bmatrix}\right] = \begin{bmatrix} w_1 & 0 \\ 0 & w_2 \end{bmatrix} .$$

Note the similarity of this model to that of Example 6.3 with $a = 1$.

The predictor updating equation is

$$2\hat{\theta}_{1t} = 2\hat{\theta}_{1t-1} + 2A_{1t}(y_t - 2\hat{\theta}_{1t-1})$$

where

$$A_{1t} = (2c_{t-1}+w_1)/(4c_{t-1}+w_1+w_2+v)$$

and

$$c_{t-1} = \text{var}(\theta_{1t} - \hat{\theta}_{1t}).$$

It is required that $0 < 2A_{1t} < 2$, which is clearly true. In fact, in the steady state

$$c = \tfrac{1}{2}\{w_1(w_2 + v)\}^{\frac{1}{2}}$$

so that

$$1-2A_1 = \frac{-w_1+w_2+v}{2\{w_1(w_2+v)\}^{\frac{1}{2}}+w_1+w_2+v} .$$

Thus the parameter can take the full range of values, $0 < 2A_1 < 2$, for when

$$(w_2 + v) \to 0, \qquad 1 - 2A_1 \to -1$$

while when

$$w_1 \to 0, \qquad 1 - 2A_1 \to 1.$$

Example 6.6 is an example of a DLM which does contain all ARIMA (0,1,1) models in the sense of the predictors being equal, and also has a diagonal covariance matrix. In

fact, the Examples 6.4, 6.5 and 6.6, without the observation error term, can be obtained from Example 6.2 by an invertible transformation $\underline{L}$ of the state vector $\underline{\theta}_t$, as described in Section 4.4. For the model (6.22),

$$\underline{L} = \begin{bmatrix} 1 & a \\ 1/a & -1 \end{bmatrix} \quad \text{where} \quad a = \left(\frac{\text{var}(w_t)}{\text{var}(v_t)} \right)^{\frac{1}{2}}$$

while for Example 6.5

$$\underline{L}_1 = \begin{bmatrix} 1 & 0 \\ -a & 1 \end{bmatrix}, \quad \text{with} \quad a = \frac{\text{cov}(w_t, v_t)}{\text{var}(w_t)} .$$

To obtain the model (6.26), we effectively applied a further transformation

$$\underline{L}_2 = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix}$$

to the model (6.24), making the total transformation from Example 6.2 to Example 6.6

$$\underline{L} = \underline{L}_2 \underline{L}_1 = \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & 1 \end{bmatrix} .$$

For the model given by (6.15), write

$$\underline{F}_o = \underline{F}, \quad \underline{G}_o = \underline{G}, \quad \underline{W}_{ot} = \underline{W}_t$$

and then apply a transformation $\underline{L}$. Using the Kalman updating equations (4.4) - (4.8), we can find expressions for $\underline{A}_t$, $\underline{C}_t$ for general invertible $\underline{L}$. We find that

$$\underline{A}_t = \underline{L}\underline{Q}_t \underline{F}_o^T / D_t$$

and

$$\underline{C}_t = \underline{L}\underline{Q}_t \{ \underline{I} - \frac{1}{D_t} \underline{F}_o^T \underline{F}_o \underline{Q}_t^T \} \underline{L}^T$$

where

$$\underline{Q}_t = \underline{G}_o \underline{L}^{-1} \underline{C}_{t-1} (\underline{L}^{-1})^T \underline{G}_o^T + \underline{W}_{ot}$$

and

$$D_t = \underline{F}_o \underline{Q}_t \underline{F}_o^T .$$

We wish to investigate the properties of the predictor, hence we wish to find $\underline{FGA}_t$ and $\underline{FGC}_t\underline{G}^T\underline{F}^T$.

$$\underline{FGA}_t = \underline{F}_o\underline{L}^{-1}\underline{LG}_o\underline{L}^{-1}\underline{LQ}_t\underline{F}_o^T/D_t$$

$$= \underline{F}_o\underline{G}_o\underline{Q}_t\underline{F}_o^T/D_t$$

and

$$\underline{FGC}_t\underline{G}^T\underline{F}^T = \underline{F}_o\underline{G}_o\underline{Q}_t\{\underline{I} - \underline{F}_o^T\underline{F}_o\underline{Q}_t^T/D_t\}\underline{G}_o^T\underline{F}_o^T$$

so the properties of the predictors are unchanged by any invertible transformation of the state vector. Clearly, a good choice of $\underline{L}$ will make $\underline{W}$ diagonal, as in Example 6.6.

6.3 The General Constant Forecast Model

We have given examples of 2 x 2 DLMs, whose predictors are equivalent in certain cases to those of the ARIMA (0,1,1) models. We now wish to investigate the general theory of such models, including those of possibly larger dimension. To do this, we shall assume that there is a steady state solution of the model, i.e. it is assumed that $v = \text{var}(v_t)$ and $\underline{W} = E[\underline{w}_t\underline{w}_t^T]$ are independent of t, and that as t becomes large, the covariance matrix $\underline{C}_t$ and hence the Kalman gain matrix $\underline{A}_t$ tend to constant matrices $\underline{C}$ and $\underline{A}$ respectively. The model is given by

$$y_t = \underline{F\theta}_t + v_t$$

$$\underline{\theta}_t = \underline{G\theta}_{t-1} + \underline{w}_t \qquad (6.29)$$

where $\underline{G}$ is of size n, and $\underline{FG}^2 = \underline{FG}$. We require an expression for the predictor $y_t(k) = \underline{FG\hat{\theta}}_t$, so we premultiply the Kalman updating equation (4.4) by $\underline{FG}$ to yield

$$\underline{FG}\hat{\theta}_t = \underline{FG}^2\hat{\theta}_{t-1} + \underline{FGA}_t(y_t - \underline{FG}\hat{\theta}_{t-1}) \qquad (6.30)$$

where it should be noted that all terms are scalars, regardless of the size n. Writing $m_t = \underline{FG}\hat{\theta}_t$ and $\mu_t = \underline{FGA}_t$ and invoking (6.4) we have

$$m_t = m_{t-1} + \mu_t(y_t - m_{t-1}) \qquad (6.31)$$

as in (6.19). In order to find the range of values taken by μ_t, it is helpful to introduce some simplifying notation. We define

$$\sigma_1^2 = \underline{FWF}^T, \quad \sigma_2^2 = \underline{FGWG}^T\underline{F}^T, \quad \rho\sigma_1\sigma_2 = \underline{FGWF}^T$$

and

$$q_t = \underline{FGC}_t\underline{G}^T\underline{F}^T = \mathrm{var}(m_t).$$

Using equations (4.4) - (4.8) and (6.4), we have

$$\underline{FGC}_t\underline{G}^T\underline{F}^T = \underline{FG}(\underline{P}_t - \underline{A}_t(\underline{FP}_t\underline{F}^T + v)\underline{A}_t^T)\underline{G}^T\underline{F}^T$$

$$= \underline{FG}\left(\underline{GC}_{t-1}\underline{G}^T + \underline{W} - \frac{\underline{P}_t\underline{F}^T\underline{FP}_t}{\underline{FP}_t\underline{F}^T + v}\right)\underline{G}^T\underline{F}^T$$

$$= \underline{FGC}_{t-1}\underline{G}^T\underline{F}^T + \underline{FGWG}^T\underline{F}^T - \frac{\underline{FG}(\underline{GC}_{t-1}\underline{G}^T + \underline{W})\underline{F}^T\underline{F}(\underline{GC}_{t-1}\underline{G}^T + \underline{W})\underline{G}^T\underline{F}^T}{\underline{FGC}_{t-1}\underline{G}^T\underline{F}^T + \underline{FWF}^T + v}.$$

Hence

$$q_t = q_{t-1} + \sigma_2^2 - \frac{(q_{t-1} + \rho\sigma_1\sigma_2)^2}{q_{t-1} + \sigma_1^2 + v}. \qquad (6.32)$$

Steady State Theory

We now move to the steady state situation, so that $q_t = q_{t-1} = q$ for all large enough t. Thus (6.32) becomes

$$q = q + \sigma_2^2 - (q + \rho\sigma_1\sigma_2)^2/(q + \sigma_1^2 + v)$$

or

$$q^2 + q(2\rho\sigma_1\sigma_2 - \sigma_2^2) + \sigma_1^2\sigma_2^2(\rho^2 - 1) - \sigma_2^2 v = 0. \qquad (6.33)$$

Lemma 6.1

If both $\underline{FG} = \underline{F}$ and $v = 0$, then the only solution to (6.33) is $q = 0$.

Proof

The expression

$$\sigma_2^2(v + \sigma_1^2(1 - \rho^2))$$

is the generalised variance of $\begin{pmatrix} v_t + \underline{Fw}_t \\ \underline{FGw}_t \end{pmatrix}$, which is positive unless $\underline{FG} = \underline{F}$ and v_t is absent. In this case, $\sigma_1^2 = \sigma_2^2$ and $\rho = 1$ and equation (6.33) reduces to $q(q + \sigma_2^2)$ for which the only admissible solution is $q = 0$. It can also be zero if $\sigma_2^2 = 0$. This implies $q = 0$ and $\underline{FGWG}^T\underline{F}^T = 0$. If $\underline{W}$ is positive definite, then it follows that $\underline{FG} = \underline{0}$ which means that $\mu = 0$ and the predictor is identically zero.

Lemma 6.2

For a non-trivial DLM, the discriminant of (6.33) is strictly positive.

Proof

The discriminant is

$$\sigma_2^2(\sigma_2^2 + 4\sigma_1^2 - 4\rho\sigma_1\sigma_2 + 4v)$$

$$= \sigma_2^2(E\left[\underline{FGw}_t - 2\underline{Fw}_t\right]^2 + 4v)$$

which is non-negative.

When this expression is zero, which happens when either

1. $\sigma_2^2 = 0$.

or

2. $\underline{FGw}_t - 2\underline{Fw}_t = 0$ and $v = 0$

then equation (6.33) has two equal roots

$$q = \tfrac{1}{2}\sigma_2(\sigma_2 - 2\rho\sigma_1).$$

1. If $\sigma_2^2 = 0$, then $q = 0$. If $\underline{W}$ is also positive definite, then $\underline{FG} = \underline{0}$, so that $\mu = 0$ and the predictor is zero with probability one. From the proof of Lemma 6.1, $\sigma_2^2 = 0$ implies that $\underline{FG} = \underline{F}$, so that $\underline{F} = \underline{0}$.

2. If $\underline{FGw}_t = 2\underline{Fw}_t$ for all $\underline{w}_t$, and $v = 0$, then $\underline{FG} = 2\underline{F}$, whence $\sigma_2^2 = 4\sigma_1^2$ and $2\rho\sigma_1\sigma_2 = 4\sigma_1^2$, hence $\rho = 1$ and $q = 0$. But from the proof of Lemma 6.1, $v = 0$ and $\rho = 1$ implies that $\underline{FG} = \underline{F}$ hence $\underline{F} = \underline{0}$.

Thus when the discriminant is zero, the observations are of the random term only, and even this may be absent.

For a non-trivial DLM, the discriminant of equation (6.33) is strictly positive, thus there are two real roots. Since '-4ac' is non-negative, the two solutions are of opposite sign. By definition, q must be positive, so we require the positive solution, that is

$$q = \tfrac{1}{2}\sigma_2(\sigma_2 - 2\rho\sigma_1) + \tfrac{1}{2}\{(2\rho\sigma_1\sigma_2 - \sigma_2^2)^2 + 4\sigma_2^2(v + (1-\rho^2)\sigma_1^2)\}^{\frac{1}{2}} \qquad (6.34)$$

$$= \tfrac{1}{2}\sigma_2\left[\sigma_2 - 2\rho\sigma_1 + \{\sigma_2^2 + 4(v - \rho\sigma_1\sigma_2 + \sigma_1^2)\}^{\frac{1}{2}}\right].$$

Now using equations (4.4) - (4.8)

$$\mu = \underline{FGA} = \underline{FGPF}^T/(\underline{FPF}^T + v)$$

$$= \underline{FG}(\underline{GCG}^T + \underline{W})\underline{F}^T/\underline{F}(\underline{GCG}^T + \underline{W})\underline{F}^T + v)$$

$$= (q + \rho\sigma_1\sigma_2)/(q + \sigma_1^2 + v)$$

or

$$1 - \mu = (\sigma_1^2 + v - \rho\sigma_1\sigma_2)/(q + \sigma_1^2 + v) \ .$$

Substituting for q from equation (6.34), we have

$$1-\mu = \frac{2(\sigma_1^2 + v - \rho\sigma_1\sigma_2)}{2(\sigma_1^2+v)+\sigma_2\{\sigma_2-2\rho\sigma_1+(\sigma_2^2+4(v-\rho\sigma_1\sigma_2+\sigma_1^2))^{\frac{1}{2}}\}} . \qquad (6.35)$$

This lies within the region $-1 < 1 - \mu < 1$.

6.4 Restrictions

It is clear from the examples in Section 6.2, that not all models satisfying $\underline{FG}^2 = \underline{FG}$ can cover the full range of the ARIMA (0,1,1) model, i.e. $0 < \underline{FGA} < 2$, but from Section 6.3, all models have the parameter lying within the region, at least in the steady state.

Theorem 6.3

If $\underline{FG} \neq \underline{F}$, then any value for μ in the range $0 < \mu < 2$ is possible.

Proof

If $\underline{FG} \neq \underline{F}$, then the expression for $1 - \mu$ is given by (6.35). If σ_2^2 is small compared with σ_1^2 and v, then $1 - \mu \to 1$, while if the variances are such that $v - \rho\sigma_1\sigma_2 + \sigma_1^2$ is close to the minimum value of $-\frac{1}{4}\sigma_2^2$, then $1 - \mu$ approaches -1. Thus $1 - \mu$ can take all values in the range $(-1,1)$ as required.

Theorem 6.4

If $\underline{FG} = \underline{F}$, then μ is restricted by $0 < \mu \leq 1$. If in addition, $v = 0$, then $\mu = 1$ and $q = 0$.

Proof

If $\underline{FG} = \underline{F}$, then $\sigma_1^2 = \sigma_2^2$ and $\rho = 1$. Hence equation (6.35) reduces to

$$1 - \mu = \frac{2v}{\sigma_1^2 + 2v + \sigma_1(\sigma_1^2 + 4v)^{\frac{1}{2}}}$$

which clearly lies between 0 and 1.

If in addition $v = 0$, then $1 - \mu = 0$ or $\mu = 1$. In this case, equation (6.34) reduces to

$$q = \tfrac{1}{2}\sigma_1(-\sigma_1 + (\sigma_1^2 + 4v)^{\frac{1}{2}}) = 0$$

In fact, the restrictions imposed by $\underline{FG} = \underline{F}$ and $v = 0$ also apply in the time dependent case. For

$$\mu_t = \underline{FGA}_t = \underline{FGP}_t\underline{F}^T/(\underline{FPF}^T + v).$$

If $\underline{FG} = \underline{F}$, then $0 < \mu_t < 1$, and from equation (6.32)

$$q_t = \frac{(q_{t-1} + \sigma_1^2)v}{q_{t-1} + \sigma_1^2 + v} .$$

Clearly, if $v = 0$, then $\mu_t = 1$ and $q_t = 0$. Thus from equation (6.31), when both $\underline{FG} = \underline{F}$ and $v = 0$, the predictor is the last observation with probability one, as in Example 6.1.

Thus if a constant forecast model is to be able to cover the full range of an ARIMA (0,1,1) model, then it must satisfy $\underline{FG} \neq \underline{F}$. However it should be pointed out that for many models with $\underline{FG} \neq \underline{F}$, it will be necessary to specify covariances between elements of the system covariance matrix as in Example 6.2 if the model is to cover the full range. Example 6.6 is a constant forecast DLM which does fulfil the complete range with a diagonal covariance matrix.

Also, when $\underline{FG} \neq \underline{F}$, there is no guarantee that μ_t will lie in the region $0 < \mu_t < 2$, even if the steady state value μ does so (see Example 6.4 with $a = \frac{1}{2}$).

In Example 6.1, we have $\underline{FG} = \underline{F}$. Clearly, when $n = 1$, the only models satisfying $fg^2 = fg$ have $g = 1$, which means that $fg = f$, or $g = 0$, which is a trivial model. So there is no scalar model which can cover the full range.

Lemma 6.5

A consequence of $\underline{FG} \neq \underline{F}$ is that $\underline{G}$ must be singular.

Proof

Consider $\underline{FG}^2 = \underline{FG}$, which is equivalent to assuming a constant forecast model. If $\underline{G}$ is non-singular then $\underline{G}^{-1}$ exists, so that

$$\underline{FG}^2\underline{G}^{-1} = \underline{FGG}^{-1}$$

or

$$\underline{FG} = \underline{F} \ .$$

Hence if $\underline{FG} \neq \underline{F}$, then $\underline{G}$ is singular.

6.5 Implications of Observability

The theory of Sections 6.3 and 6.4 was derived for general n, hence it would seem that models of any dimension greater than one, satisfying $\underline{FG}^2 = \underline{FG}$ and $\underline{FG} \neq \underline{F}$ are appropriate. We shall show that there is also an upper limit on the dimension of such models.

In Section 5.3, the observability of a DLM was defined as the ability to estimate the state $\underline{\theta}_t$ from the past observations y_{t-j} $j \geq 0$. But Theorem 5.6 due to Kalman (1963a) means that observability together with positive definite $\underline{W}$ is sufficient condition for the equilibrium state to exist. Since

the theory depends on the existence of the steady state, it is desirable that all our models are observable. The criterion for observability given in that section is that the matrix

$$\left[\underline{F}^T(\underline{FG})^T(\underline{FG}^2)^T\ldots(\underline{FG}^{n-1})^T\right] \tag{6.36}$$

should be of rank n, where n is the dimension of $\underline{G}$. Hence we have

Lemma 6.6

If the DLM is observable, then $n = 1$ or $n = 2$.

Proof.

Since $\underline{FG}^2 = \underline{FG}$, it is clear from (6.36) that for $n \geq 3$, the matrix has rank less than n, hence result.

We have shown in Section 6.4 that models with $n = 1$ cannot cover the full range.

When $n = 2$, we see from (6.36) that the requirement $\underline{FG} \neq \underline{F}$ is equivalent to stipulating that the model be observable. Example 6.3 is not observable, and hence may not have an equilibrium solution. Examples 6.2 and 6.4 - 6.6 are all observable, since they all satisfy $\underline{FG} \neq \underline{F}$.

CHAPTER 7

THE POLYNOMIAL MODEL

7.1 Representation of the Model

We now consider the more general case, where the forecast function $\{y_t(k);\ k = 1,2,\ldots\}$ is a polynomial of degree $d - 1$, that is

$$y_t(k) = a_o + a_1 k + a_2 k^2 + \ldots + a_{d-1} k^{d-1} \qquad k \geq 1 \qquad (7.1)$$

where $a_{d-1} \neq 0$.

It is convenient to represent the polynomial (7.1) in an alternative formulation. For any given function $f(x)$, we define the backward shift operator ∇ by $\nabla f(x) = f(x) - f(x - 1)$. It is well known that $\nabla^{n+1} f(x) \equiv 0$ if and only if $f(x)$ is a polynomial of degree n defined on the integers. Using this result, the representation (7.1) is equivalent to

$$\nabla^d y_t(k) \equiv 0$$

or

$$\sum_{i=o}^{d} \binom{d}{i} (-1)^i y_t(k-i) \equiv 0 \qquad k \geq d + 1 . \qquad (7.2)$$

Since the predictor $y_t(k)$ derived in Section 4.3 is given by

$$y_t(k) = \underline{F}\underline{G}^k \hat{\underline{\theta}}_t \qquad \text{for all} \quad k \geq 1$$

equation (7.2) becomes

$$\sum_{i=o}^{d} \binom{d}{i} (-1)^i \underline{F}\underline{G}^{k-i} \hat{\underline{\theta}}_t \equiv 0 \qquad k \geq d + 1 .$$

This must hold for all values of $\hat{\underline{\theta}}_t$, so we have

$$\sum_{i=o}^{d} \binom{d}{i} (-1)^i \underline{FG}^{k-i} = \underline{O} \qquad k \geq d + 1$$

or

$$\underline{FG}^j (\underline{G}-\underline{I})^d = \underline{O} \qquad j \geq 1 .$$

In particular,

$$\underline{FG}(\underline{G}-\underline{I})^d = \underline{O} . \tag{7.3}$$

Notice that because $a_{d-1} \neq 0$, then $\nabla^{d-1} y_t(k) \not\equiv 0$, or by the above argument,

$$\underline{FG}(\underline{G}-\underline{I})^{d-1} \not\equiv \underline{O} . \tag{7.4}$$

It is clear that if we can find $y_t(1), \ldots y_t(d)$, then we can find $y_t(k)$ for $k \geq d + 1$ recursively from equation (7.2). For some purposes, it may be easier to find some invertible transformation of these quantities. In this discussion, we shall find it convenient to consider the quantities $\underline{FG}\hat{\theta}_t$, $\underline{FG}(\underline{G}-\underline{I})\hat{\theta}_t, \ldots, \underline{FG}(\underline{G}-\underline{I})^{d-1}\hat{\theta}_t$. We define

$$\underline{R}_d = \begin{bmatrix} \underline{FG} \\ \underline{FG}(\underline{G}-\underline{I}) \\ \underline{FG}(\underline{G}-\underline{I})^2 \\ \vdots \\ \underline{FG}(\underline{G}-\underline{I})^{d-2} \\ \underline{FG}(\underline{G}-\underline{I})^{d-1} \end{bmatrix} . \tag{7.5}$$

Since $\underline{FG}(\underline{G}-\underline{I})^r = \sum_{i=o}^{r} \binom{r}{i} (-1)^{r-i} \underline{FG}^{i+1}$, we have

$$\underline{R}_d = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & & 0 \\ -1 & 3 & -3 & 1 & & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & \vdots & 1 & 0 \\ (-1)^{d-1} & (-1)^{d-2}(d-1) & \cdots & \cdots & -(d-1) & 1 \end{bmatrix} \begin{bmatrix} \underline{FG} \\ \underline{FG}^2 \\ \underline{FG}^3 \\ \vdots \\ \vdots \\ \vdots \\ \underline{FG}^{d-1} \\ \underline{FG}^d \end{bmatrix} \qquad (7.6)$$

The (i,j)th term of this transformation matrix is

$$\begin{cases} \binom{i-1}{j-1}(-1)^{i-j} & i \geq j \\ 0 & i < j \, . \end{cases}$$

The matrix is invertible, thus $\underline{R}_d\hat{\underline{\theta}}_t$ is an invertible transformation of $y_t(1),\ldots y_t(d)$. None of the rows of the $(d \times n)$ matrix $\underline{R}_d$ can be zero, because by (7.4), this would imply that the forecast function is a polynomial of degree less than $d - 1$.

Lemma 7.1

The matrix $\underline{R}_d$ has rank d.

Proof

Suppose $\underline{R}_d$ has rank less than d. Then the rows of $\underline{R}_d$ are linearly dependent, and we can find coefficients a_i, not all zero, such that

$$\sum_{i=1}^{d} a_i \underline{FG}(\underline{G}-\underline{I})^{i-1} = \underline{O} \, .$$

Suppose that k $(1 \leq k \leq d - 1)$ is the smallest integer such that $a_k \neq 0$. Then

$$\sum_{i=k}^{d} a_i \underline{FG}(\underline{G}-\underline{I})^{i-1} = \underline{O} \, .$$

Postmultiply by $(\underline{G}-\underline{I})^{d-k}$, which from (7.4) is not zero, to yield

$$\sum_{i=k}^{d} a_i \underline{FG}(\underline{G}-\underline{I})^{d+i-k-1} = \underline{O} \ .$$

Using (7.3), this reduces to

$$a_k \underline{FG}(\underline{G}-\underline{I})^{d-1} = \underline{O} \ .$$

But $\underline{FG}(\underline{G}-\underline{I})^{d-1} \neq \underline{O}$ by (7.4) and $a_k \neq 0$ by construction. This contradiction shows that the rows of $\underline{R}_d$ are not linearly dependent, hence the matrix has rank d.

We now premultiply the Kalman updating equation (4.4) by $\underline{R}_d$ to give

$$\underline{R}_d\hat{\underline{\theta}}_t = \underline{R}_d\underline{G}\hat{\underline{\theta}}_{t-1} + \underline{R}_d\underline{A}_t(y_t - \underline{FG}\hat{\underline{\theta}}_{t-1}) \ . \qquad (7.7)$$

Using the fact that

$$\underline{FG}^2(\underline{G}-\underline{I})^{i-1} = \underline{FG}(\underline{G}-\underline{I})^{i} + \underline{FG}(\underline{G}-\underline{I})^{i-1}$$

and equation (7.3), we find that

$$\underline{R}_d\underline{G} = \begin{bmatrix} \underline{FG}^2 \\ \underline{FG}^2(\underline{G}-\underline{I}) \\ \vdots \\ \underline{FG}^2(\underline{G}-\underline{I})^{d-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & & \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \underline{FG} \\ \underline{FG}(\underline{G}-\underline{I}) \\ \vdots \\ \underline{FG}(\underline{G}-\underline{I})^{d-1} \end{bmatrix} . \qquad (7.8)$$

Writing

$$m_{it} = \underline{FG}(\underline{G}-\underline{I})^{i-1}\hat{\underline{\theta}}_t, \quad \mu_{it} = \underline{FG}(\underline{G}-\underline{I})^{i-1}\underline{A}_t \quad i = 1,\ldots d \qquad (7.9)$$

and

$$\underline{M}_t = (m_{1t},\ldots m_{dt})^T, \quad \underline{\mu}_t = (\mu_{1t},\ldots \mu_{dt})^T,$$

equation (7.7) becomes

$$\underline{M}_t = \underline{K}\underline{M}_{t-1} + \underline{\mu}_t(y_t - m_{1t-1}) \qquad (7.10)$$

where

$$\underline{K} = \begin{cases} 1 & i = j, \ \ i = j-1 \\ 0 & \text{elsewhere} \end{cases}$$

which is exactly the same formulation as equation (3.2)

of Godolphin and Harrison (1975) (It should be noted that the definition of $\underline{K}$ given there is incorrect, since the transpose of $\underline{K}$ is given, although $\underline{K}$ itself is derived). We can also show

Lemma 7.2

$$y_t(k) = \underline{FG}^k\hat{\underline{\theta}}_t = \sum_{j=1}^{d} \binom{k-1}{j-1} m_{jt} \quad k \geq 1 . \tag{7.11}$$

This is equation (3.1) of Godolphin and Harrison (1975) and leads to the conjecture that m_{it} defined in equation (7.9) is the same as the ψ_{it} of Godolphin and Harrison (1975).

Proof

$$\sum_{j=1}^{d} \binom{k-1}{j-1} m_{jt} = \sum_{j=1}^{d} \binom{k-1}{j-1} \underline{FG}(\underline{G}-\underline{I})^{j-1}\hat{\underline{\theta}}_t$$

$$= \sum_{j=1}^{d} \binom{k-1}{j-1} \underline{FG} \sum_{i=o}^{j-1} \binom{j-1}{i} (-1)^{j-1-i} \underline{G}^i \hat{\underline{\theta}}_t$$

$$= \sum_{i=o}^{d-1} \underline{FG}^{i+1} \binom{k-1}{i} \sum_{r=o}^{d-i-1} \binom{k-1-i}{k-1-i-r} (-1)^r \hat{\underline{\theta}}_t .$$

When $k = d$

$$\sum_{r=o}^{d-i-1} \binom{d-1-i}{d-1-i-r} (-1)^r = \begin{cases} o & d-1 > i \\ 1 & d-1 = i \end{cases}$$

so that

$$\sum_{j=1}^{d} \binom{d-1}{j-1} m_{jt} = \underline{FG}^d \hat{\underline{\theta}}_t = y_t(d) .$$

When $k < d$, $\binom{k-1}{j-1} = 0$ for $j > k$, so that

$$\sum_{j=1}^{d} \binom{k-1}{j-1} m_{jt} \text{ becomes}$$

$$\sum_{i=0}^{k-1} \underline{FG}^{i+1} \binom{k-1}{i} \sum_{r=0}^{k-i-1} \binom{k-1-i}{k-1-i-r} (-1)^r \hat{\underline{\theta}}_t$$

which by the above argument is $\underline{FG}^k \hat{\underline{\theta}}_t = y_t(k)$. Thus (7.11) is proved for $1 \leq k \leq d$, hence by (7.2) it is true for all $k \geq 1$.

Comparison of equations (7.10) and (7.11) with the work of Godolphin and Harrison (1975) suggests we are dealing with an ARIMA (0,d,q) process, where $q \leq d$. Before we can say this however, we need to know about the range of values that can be taken by $\underline{\mu}_t$ in this model.

7.2 Steady State Theory

We assume steady state conditions, since it is only in the equilibrium state that comparisons with ARIMA processes are applicable. Thus we assume $\underline{\mu}_t = \underline{\mu}$ for all t, however $\underline{M}_t$ and y_t continue to change with t. Hence we have from equation (7.10)

$$\underline{M}_t = \underline{KM}_{t-1} + \underline{\mu}(y_t - m_{1t-1}) \quad . \qquad (7.12)$$

We can now apply z-transform theory, and the concept of stability, as discussed in Section 5.1. Taking z-transforms of equation (7.12), we have

$$\underline{M}(z) = z^{-1}\underline{KM}(z) + \underline{\mu}(Y(z) - z^{-1}M_1(z)) \quad . \qquad (7.13)$$

This set of d equations must be solved to find $(M_i(z))/Y(z)$ in terms of $\mu_1, \ldots \mu_d$. For a stable system, the characteristic polynomial of the filter, which is the denominator of $M_i(z)/Y(z)$ must have all roots less than one in modulus. Writing out (7.13) in detail, we have

$$M_i(z) = z^{-1}(M_i(z) + M_{i+1}(z)) + \mu_i(Y(z) - z^{-1}M_1(z))$$

$$i = 1,\ldots d-1$$

$$M_d(z) = z^{-1}M_d(z) + \mu_d(Y(z) - z^{-1}M_1(z)). \qquad (7.14)$$

Lemma 7.3

Equations (7.13) satisfy

$$M_{d-r}(z) = \sum_{j=o}^{r} \mu_{d-r+j} \frac{(zY(z)-M_1(z))}{(z-1)^{j+1}} \quad r = 0,1,\ldots d-1 . \qquad (7.15)$$

Proof

From (7.14)

$$M_d(z) = \mu_d \frac{(zY(z)-M_1(z))}{z-1}$$

which is (7.15) with r = 0.

We now proceed by induction. Suppose (7.15) is true for a particular value of r. Then using (7.14)

$$M_{d-(r+1)}(z) = z^{-1}(M_{d-r-1}(z) + M_{d-r}(z)) + \mu_{d-r-1}(Y(z)-z^{-1}M_1(z)).$$

But $M_{d-r}(z)$ is given by (7.15), hence

$$(z-1)M_{d-r-1}(z) = \sum_{j=o}^{r} \mu_{d-r+j} \frac{(zY(z)-M_1(z))}{(z-1)^{j+1}} + \mu_{d-r-1}(zY(z)-M_1(z))$$

$$= \sum_{j=-1}^{r} \mu_{d-r+j} \frac{(zY(z)-M_1(z))}{(z-1)^{j+1}} .$$

Thus

$$M_{d-r-1}(z) = \sum_{i=o}^{r+1} \mu_{d-r-1+i} \frac{(zY(z)-M_1(z))}{(z-1)^{i+1}}$$

which is (7.15) with r replaced by r + 1. Thus the induction is proved and (7.15) holds for r = 0,1,...d-1.

Lemma 7.4

The characteristic polynomial of the filter is given by

$$\beta(z) = \sum_{i=o}^{d} \beta_i z^{d-i} \qquad (7.16)$$

where

$$\beta_o = 1, \quad \beta_i = \binom{d}{i}(-1)^i + \sum_{j=1}^{i} \mu_j \binom{d-j}{i-j}(-1)^{i-j} . \qquad (7.17)$$

Proof

The characteristic polynomial for the model is the denominator of $M_i(z)/Y(z)$. Putting $r = d-1$ in equation (7.15), we have

$$M_1(z) = \sum_{j=o}^{d-1} \mu_{j+1} \frac{(zY(z)-M_1(z))}{(z-1)^{j+1}} .$$

Collecting terms

$$M_1(z)\left[1 + \sum_{j=o}^{d-1} \mu_{j+1}/(z-1)^{j+1}\right] = zY(z) \sum_{j=o}^{d-1} \mu_{j+1}/(z-1)^{j+1}$$

or

$$M_1(z)\left[(z-1)^d + \sum_{j=o}^{d-1} \mu_{j+1}(z-1)^{d-j-1}\right] = zY(z) \sum_{j=o}^{d-1} \mu_{j+1}(z-1)^{d-j-1}$$

so that

$$\frac{M_1(z)}{Y(z)} = \frac{z \sum_{j=o}^{d-1} \mu_{j+1}(z-1)^{d-j-1}}{\left[(z-1)^d + \sum_{j=1}^{d} \mu_j (z-1)^{d-j}\right]} .$$

Hence the characteristic polynomial $\beta(z)$ is given by

$$\beta(z) = (z-1)^d + \sum_{j=1}^{d} \mu_j (z-1)^{d-j}$$

$$= \sum_{i=o}^{d} \binom{d}{i}(-1)^i z^{d-i} + \sum_{j=1}^{d} \mu_j \sum_{i=o}^{d-j} \binom{d-j}{i}(-1)^i z^{d-j-i}$$

$$= \sum_{i=o}^{d} \binom{d}{i}(-1)^i z^{d-i} + \sum_{j=1}^{d} \mu_j \sum_{k=j}^{d} \binom{d-j}{k-j}(-1)^{k-j} z^{d-k}$$

$$= \sum_{i=o}^{d} \binom{d}{i}(-1)^i z^{d-i} + \sum_{k=1}^{d} z^{d-k} \sum_{j=1}^{k} \mu_j \binom{d-j}{k-j}(-1)^{k-j} .$$

Thus the coefficient of z^d is 1, and for $1 \leq i \leq d$, the coefficient of z^{d-i} is

$$\beta_i = \binom{d}{i}(-1)^i + \sum_{j=1}^{i} \mu_j \binom{d-j}{i-j}(-1)^{i-j}.$$

This expression for the β_i is not always helpful for the derivation of results. An alternative formulation is derived in the following lemma.

Lemma 7.5

β_i can be expressed as

$$\beta_i = \binom{d}{i}(-1)^i + \sum_{j=0}^{i-1} \binom{d}{j}(-1)^j \underline{FG}^{i-j}\underline{A} \qquad (7.18)$$

where $\underline{A}$ is the limit of $\underline{A}_t$ in the steady state.

Proof

We need to show that

$\sum_{j=1}^{i} \mu_j \binom{d-j}{i-j}(-1)^{i-j}$ is equal to $\sum_{j=0}^{i-1} \binom{d}{j}(-1)^j \underline{FG}^{i-j}\underline{A}$.

First substitute $\mu_j = \underline{FG}(\underline{G}-\underline{I})^{j-1}\underline{A}$ into the left hand side. Then this becomes

$$\sum_{j=1}^{i} \binom{d-j}{i-j}(-1)^{i-j}\underline{FG} \sum_{k=0}^{j-1} \binom{j-1}{k}(-1)^k \underline{G}^{j-1-k}\underline{A}$$

$$= \sum_{j=1}^{i} \sum_{k=0}^{j-1} \binom{d-j}{i-j}\binom{j-1}{k}(-1)^{i-j+k}\underline{FG}^{j-k}\underline{A}$$

$$= \sum_{k=0}^{i-1} \sum_{j=k+1}^{i} \binom{d-j}{i-j}\binom{j-1}{k}(-1)^{i-j+k}\underline{FG}^{j-k}\underline{A}$$

$$= \sum_{k=0}^{i-1} \sum_{r=1}^{i-k} \binom{d-r-k}{d-i}\binom{r+k-1}{k}(-1)^{i-r}\underline{FG}^r\underline{A}$$

$$= \sum_{r=1}^{i} (-1)^{i-r}\underline{FG}^r\underline{A} \sum_{k=0}^{i-r} \binom{d-r-k}{i-r-k}\binom{r+k-1}{k}.$$

A well known identity from combinatorial analysis is

$$\sum_{k=0}^{j} \binom{a+j-k-1}{j-k}\binom{b+k-1}{k} = \binom{a+b+j-1}{j}$$

(Feller, 1968, p.65).

Putting $b = r$, $j = i - r$ and $a = d - i + 1$ in this identity, we have the inner summation above. Hence

$$\sum_{j=1}^{i} \mu_j \binom{d-j}{i-j}(-1)^{i-j} = \sum_{r=1}^{i} (-1)^{i-r}\underline{FG}^r\underline{A}\binom{d}{i-r}$$

$$= \sum_{j=0}^{i-1} (-1)^j \binom{d}{j}\underline{FG}^{i-j}\underline{A}$$

as required, thus equation (7.18) holds. Notice that (7.18) is exactly the definition of β_i given by Godolphin and Harrison (1975) (see equation (3.18), where the α_j of Theorem 3.1 are given by $\alpha_j = \underline{FG}^j\underline{A}$ for $j = 1,\ldots,d$. Collecting all these results together we have the following

Theorem 7.6

Suppose the DLM satisfies $\underline{FG}(\underline{G}-\underline{I})^d = \underline{0}$, $\underline{FG}(\underline{G}-\underline{I})^{d-1} \neq \underline{0}$. Then in the steady state situation, the predictor $y_t(k) = \underline{FG}^k\hat{\underline{\theta}}_t$ is identical to the k-step ahead predictor of an ARIMA (0,d,q) process ($q \leq d$) if and only if the model's estimation scheme is stable in the steady state.

Proof

The proof of this theorem follows directly from Theorem 3.1. Equation (3.17) is, in the notation of the DLM

$$\underline{FG}^k\hat{\underline{\theta}}_t = \underline{FG}^{k+1}\hat{\underline{\theta}}_{t-1} + \underline{FG}^k\underline{A}(y_t - \underline{FG}\hat{\underline{\theta}}_{t-1}),$$

which is the Kalman updating equation (4.4) premultiplied by $\underline{FG}^k$, which clearly holds. If $1 + \beta_1 z + \ldots + \beta_d z^d$ has

all zeros strictly outside the unit circle, then $z^d + \beta_1 z^{d-1} + \ldots + \beta_d$ has all zeros strictly inside the unit circle, which is the stability condition for the DLM, with β_i given by (7.18).

Thus we see that in general, the stability conditions for the estimation scheme of the DLMs described in this chapter are equivalent to the invertibility conditions of the ARIMA (0,d,q) ($q \leq d$) models.

In the following example of a polynomial model of degree one, we shall investigate the range of values taken by β_1, β_2 and compare this to the invertibility region of an ARIMA (0,2,2) process.

Example 7.1

Consider the model given by

$$Y_t = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} \underline{\theta}_t + \varepsilon_t$$

$$\underline{\theta}_t = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \underline{\theta}_{t-1} + \underline{w}_t$$

where

$$\underline{W} = \begin{bmatrix} w_1 & & & \\ & w_2 & & \\ & & w_3 & \\ & & & w_4 \end{bmatrix}$$ is diagonal

and

$$v = \text{var}(\varepsilon_t).$$

This satisfies

$$\underline{FG}(\underline{G}-\underline{I})^2 = \underline{0}, \quad \underline{FG}(\underline{G}-\underline{I}) \neq \underline{0}, \quad \underline{F}(\underline{G}-\underline{I})^2 \neq \underline{0}$$

hence is a polynomial model of degree one.

From (5.19), the observability criterion requires that $\underline{F}$, $\underline{FG}$, $\underline{FG}^2$, $\underline{FG}^3$ are linearly independent. But $\underline{FG}(\underline{G}-\underline{I})^2 = 0$,

so that this model is not observable, hence we do not know whether or not an equilibrium solution exists. We shall assume its existence, and derive the values of β_2, $1 + \beta_1 + \beta_2$, $1 - \beta_1 + \beta_2$.

Applying the Kalman updating equations (4.4) - (4.8) we find that $\underline{A}_t = \left[A_{1t}\ A_{2t}\ A_{3t}\ A_{4t}\right]^T$ is given by

$$A_{1t} = (2p_{t-1} + w_1)/D_t$$

$$A_{2t} = (2p_{t-1} + w_2)/D_t$$

$$A_{3t} = (2q_{t-1} + w_3)/D_t$$

$$A_{4t} = (2q_{t-1} + w_4)/D_t$$

where

$$p_t = \mathrm{var}(\theta_{1t}-\hat{\theta}_{1t})+3\mathrm{cov}(\theta_{1t}-\hat{\theta}_{1t},\theta_{3t}-\hat{\theta}_{3t})+2\mathrm{var}(\theta_{3t}-\hat{\theta}_{3t})$$

$$q_t = \mathrm{cov}(\theta_{1t}-\hat{\theta}_{1t},\theta_{3t}-\hat{\theta}_{3t})+2\mathrm{var}(\theta_{3t}-\hat{\theta}_{3t})$$

and

$$D_t = 4(p_{t-1}+q_{t-1})+w_1+w_2+w_3+w_4+v$$

$$= 4\mathrm{var}\left[(\theta_{1t-1}-\hat{\theta}_{1t-1})+2(\theta_{3t-1}-\hat{\theta}_{3t-1})\right]+w_1+w_2+w_3+w_4+v\ .$$

From equation (7.18)

$$\beta_1 = -2+\underline{FGA} = -2+2A_1+4A_3$$

$$\beta_2 = 1-2\underline{FGA}+\underline{FG}^2\underline{A} = 1-2(A_1+A_3)$$

where A_1, A_3 are the steady state values of A_{1t}, A_{3t}. We wish to show

(i) $|\beta_2| < 1$

(ii) $1 + \beta_1 + \beta_2 > 0$

(iii) $1 - \beta_1 + \beta_2 > 0$.

We shall use the time dependent values of A_{1t}, A_{3t}, for if the inequalities are satisfied for all t, they are

satisfied in the steady state .

(i) $\beta_2 = 1 - 2A_{1t} - 2A_{3t}$

$$= \frac{4(p_{t-1}+q_{t-1})+w_1+w_2+w_3+w_4+v-2(2p_{t-1}+w_1)-2(2q_{t-1}+w_3)}{D_t}$$

$$= \frac{-w_1+w_2-w_3+w_4+v}{4(p_{t-1}+q_{t-1})+w_1+w_2+w_3+w_4+v}$$

which clearly lies between -1 and 1.

(ii) $1+\beta_1+\beta_2 = 1-2+2A_{1t}+4A_{3t}+1-2A_{1t}-2A_{3t}$

$$= 2A_{3t}$$

$$= 2(2q_{t-1}+w_3)/D_t \ .$$

If $\mathrm{cov}(\theta_{1t}-\hat{\theta}_{1t}, \theta_{3t}-\hat{\theta}_{3t})+\mathrm{var}(\theta_{3t}-\hat{\theta}_{3t})$ is positive, then q_t is positive, hence $1+\beta_1+\beta_2 > 0$.

(iii) $1-\beta_1+\beta_2 = 1+2-2A_{1t}-4A_{3t}+1-2A_{1t}-2A_{3t}$

$$= 4-4A_{1t}-6A_{3t}$$

$$= 2(2-2A_{1t}-3A_{3t}) \ .$$

$$D_t(2-2A_{1t}-3A_{3t}) = 2\left[4(p_{t-1}+q_{t-1})+w_1+w_2+w_3+w_4+v\right] - 2(2p_{t-1}+w_1)-3(2q_{t-1}+w_3)$$

$$= 4p_{t-1}+2q_{t-1}+2w_2-w_3+2w_4+2v.$$

If $\mathrm{cov}(\theta_{1t}-\hat{\theta}_{1t}, \theta_{3t}-\hat{\theta}_{3t})+\mathrm{var}(\theta_{3t}-\hat{\theta}_{3t})$ is positive, as required for (ii), and also

$$2(w_2+w_4+v) > w_3$$

then $1-\beta_1+\beta_2 > 0$.

Even if these conditions are satisfied, so that the parameters lie within the invertibility region, it is not

clear from the above whether or not the whole region can be covered. To examine this, we consider the second backward differenced series

$$\nabla^2 y_t = w_{1t} - w_{1t-2} + w_{3t} + 2w_{3t-1} - w_{3t-2} + \nabla^2 (w_{2t} + w_{4t} + \varepsilon_t)$$

which has the same correlation structure as the ARIMA (0,2,2) model

$$\nabla^2 y_t = a_t + \beta_1 a_{t-1} + \beta_2 a_{t-2} \quad .$$

Multiply $\nabla^2 y_t$ in turn by $\nabla^2 y_t$, $\nabla^2 y_{t-1}$ and $\nabla^2 y_{t-2}$, and take expectations, then equating the results obtained from the two equations we find that

$$E\left[(\nabla^2 y_t)^2\right] = 2w_1 + 6w_3 + 6(w_2 + w_4 + v) = (1 + \beta_1^2 + \beta_2^2)\,\mathrm{var}(a)$$

$$E\left[(\nabla^2 y_t)(\nabla^2 y_{t-1})\right] = -4(w_2 + w_4 + v) = \beta_1(1 + \beta_2)\,\mathrm{var}(a)$$

$$E\left[(\nabla^2 y_t)(\nabla^2 y_{t-2})\right] = -w_1 + w_2 - w_3 + w_4 + v = \beta_2\,\mathrm{var}(a) \quad .$$

Since we know $\beta_2 > -1$, the second equation implies $\beta_1 < 0$.

Solving these equations

$$w_2 + w_4 + v = -\beta_1(1 + \beta_2)\,\mathrm{var}(a)/4$$

$$w_3 = (1 + \beta_1 + \beta_2)^2 \mathrm{var}(a)/4$$

$$w_1 = \left[-(1 + \beta_1 + \beta_2)^2 - \beta_1(1 + \beta_2) - 4\beta_2\right]\mathrm{var}(a)/4 \quad .$$

Since this expression for w_1 should be positive, we have

$$\beta_2^2 + \beta_2(3\beta_1 + 6) + \beta_1^2 + 3\beta_1 + 1 < 0 \quad .$$

Thus the range covered by this model is given by the shaded region in Figure 7.1.

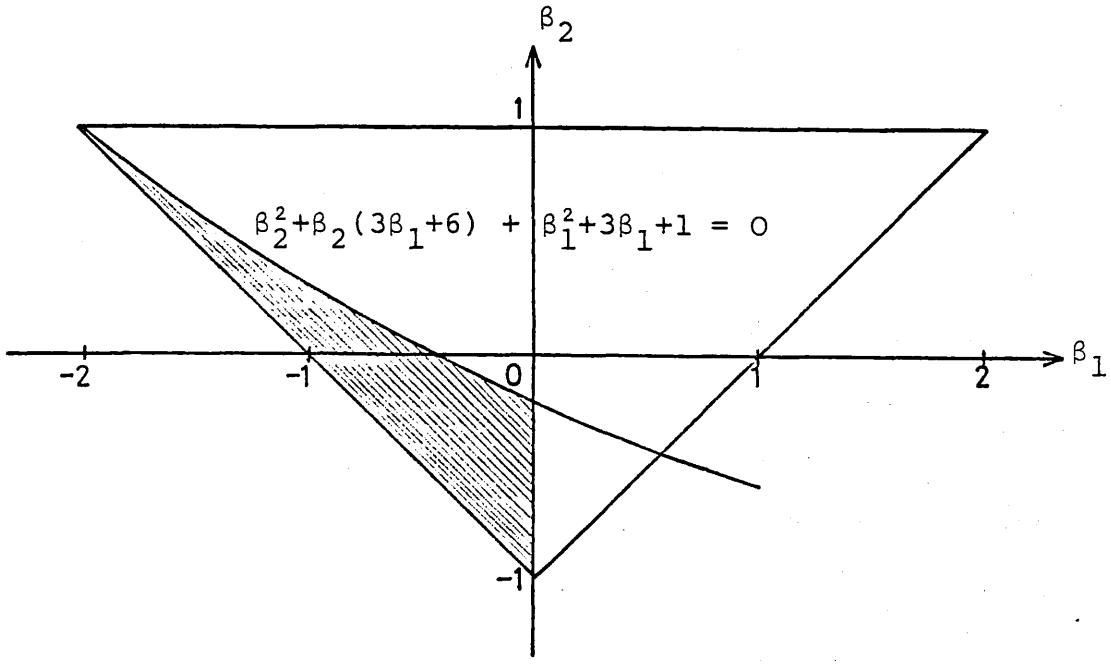


FIGURE 7.1 :STABILITY REGION FOR EXAMPLE 7.1

7.3 Size and Structure of $\underline{G}$

Intuitively, we expect a model with a polynomial forecast function of degree d-1 to have d system variables, that is, we expect $\underline{G}$ to have dimension d. The results for the constant forecast model in the previous chapter suggest that the dimension of the system vector $\underline{\theta}_t$ should be increased.

Example 7.2

We examine the 'linear growth' model used by Harrison and Stevens (1976)

$$\begin{aligned} y_t &= u_t + \varepsilon_t \\ u_t &= u_{t-1} + \beta_t + \delta u_t \\ \beta_t &= \beta_{t-1} + \delta\beta_t \ . \end{aligned} \tag{7.19}$$

That is, $\underline{F} = \begin{bmatrix} 1 & 0 \end{bmatrix}$, $\underline{G} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and

$$\underline{W} = \begin{bmatrix} \text{var}(\delta u_t+\delta\beta_t) & \text{var}(\delta\beta_t)+\text{cov}(\delta u_t,\delta\beta_t) \\ \text{var}(\delta\beta_t)+\text{cov}(\delta u_t,\delta\beta_t) & \text{var}(\delta\beta_t) \end{bmatrix} .$$

The model satisfies $\underline{FG}(\underline{G}-\underline{I})^2 = \underline{O}$, $\underline{FG}(\underline{G}-\underline{I}) \neq \underline{O}$. From equation (5.19), the criterion for observability is that $\text{rank}\left[\underline{F}^T \ \underline{FG}^T\right] = 2$. Since $\underline{FG} = \begin{bmatrix} 1 & 1 \end{bmatrix} \neq \alpha\underline{F}$ for any α, the model is observable. Hence if we also assume $\underline{W}$ is positive definite, the corollary to Theorem 5.6 implies that this model will converge to an equilibrium state, thus the theory of Section 7.2 is applicable. We shall assume that the model is stable in the steady state, then by Theorem 7.6, its predictor is identical to the predictor of an ARIMA (0,2,2) process. In fact, we can write the model (7.19) as an ARIMA (0,2,2) process as follows:

$$\begin{aligned} \nabla^2 y_t &= \nabla\beta_t+\nabla\delta u_t+\nabla^2\varepsilon_t \\ &= \delta\beta_t+\nabla\delta u_t+\nabla^2\varepsilon_t \end{aligned} \tag{7.20}$$

which has the same autocorrelation structure as the ARIMA (0,2,2) model

$$\nabla^2 y_t = a_t+\beta_1 a_{t-1}+\beta_2 a_{t-2} . \tag{7.21}$$

Multiply $\nabla^2 y_t$ in turn by $\nabla^2 y_t$, $\nabla^2 y_{t-1}$, $\nabla^2 y_{t-2}$ and take expectations, then equating the results obtained from equations (7.20) and (7.21), we have

$$\begin{aligned} E\left[(\nabla^2 y_t)^2\right] &= (1+\beta_1^2+\beta_2^2)\text{var}(a) \\ &= \text{var}(\delta\beta)+2\text{var}(\delta u)+2\text{cov}(\delta u,\delta\beta)+6\text{var}(\varepsilon) \end{aligned}$$

$$E\left[(\nabla^2 y_t)(\nabla^2 y_{t-1})\right] = \beta_1(1+\beta_2)\,var(a)$$
$$= -cov(\delta u, \delta\beta) - var(\delta u) - 4var(\varepsilon)$$
$$E\left[(\nabla^2 y_t)(\nabla^2 y_{t-2})\right] = \beta_2 var(a) = var(\varepsilon),$$

where it is assumed that the variances are independent of t, e.g. $var(\delta\beta_t) = var(\delta\beta)$ for all t.

Solving these three equations

$$var(\varepsilon) = \beta_2 var(a)$$

$$cov(\delta u, \delta\beta) + var(\delta u) = -(\beta_1(1+\beta_2) - 4\beta_2)\, var(a)$$

$$var(\delta\beta) = (1+\beta_1^2+\beta_2^2+2\beta_1(1+\beta_2)+8\beta_2-6\beta_2)\,var(a)$$

$$= (1+\beta_1+\beta_2)^2 var(a)\ .$$

Obviously, these results imply $\beta_2 \geq 0$. If, in addition, we require $cov(\delta u, \delta\beta) = 0$, then it follows that

$$\beta_1(1+\beta_2) - 4\beta_2 \leq 0$$

or

$$\beta_1 \leq -4\beta_2/(1+\beta_2). \qquad (7.22)$$

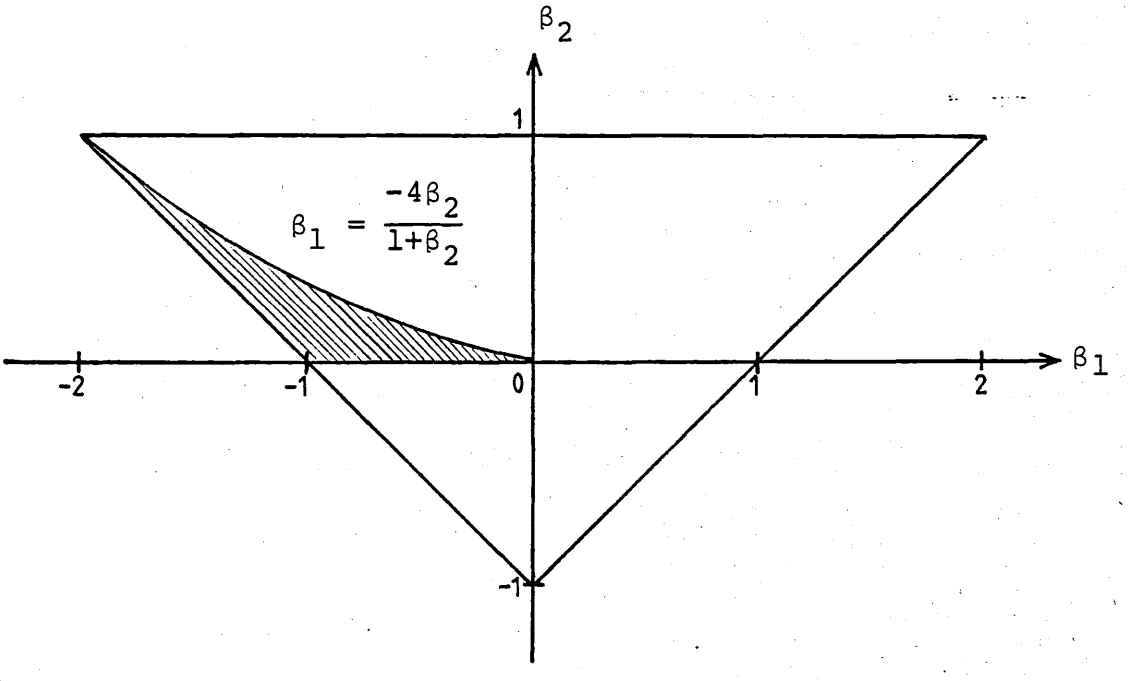


FIGURE 7.2: STABILITY REGION FOR THE LINEAR GROWTH MODEL

The inequality (7.22) is an even greater restriction than that noted by Godolphin (1976b) in his discussion of Harrison and Stevens paper. In fact, this restriction also holds whenever $var(\delta\mu) \geq var(\delta\beta)$, and this includes the other case of interest, namely $\underline{W}$ diagonal. For

$$var(\delta\mu) \geq var(\delta\beta)$$

implies that

$$var(\delta\mu) + cov(\delta\mu,\delta\beta) \geq var(\delta\beta) + cov(\delta\mu,\delta\beta).$$

But

$$\{var(\delta\mu)+cov(\delta\mu,\delta\beta)\}+\{cov(\delta\mu,\delta\beta)+var(\delta\beta)\} \geq 0$$

so that

$$var(\delta\mu)+cov(\delta\mu,\delta\beta) \geq 0$$

from which the restriction follows.

Thus we see that the stability region for this model, given by the shaded region in Figure 7.2, is rather restricted, in fact more restricted than that for Example 7.1. By analogy with the constant forecast model, it is reasonable to suppose that the dimension of $\underline{G}$ should be increased. We confirm this conjecture by investigating the dimension and rank of several matrices.

Lemma 7.7

The dimension of $\underline{G}$ is not less than d.

Proof

This is a direct consequence of Lemma 7.1, where the $d \times n$ matrix $\underline{R}_d$ was shown to have rank d. This implies that n, the dimension of $\underline{G}$, is greater than or equal to d.

This lemma has the intuitively sensible interpretation that any model which has a forecast function which is a polynomial of degree $d - 1$ must have at least d system variables.

To ensure that the model converges to an equilibrium state, and hence that the theory of Section 7.2 is applicable, we require that the model is observable, i.e. we require

$$\left[\underline{F},\ \underline{FG},\ldots\underline{FG}^{n-1}\right] \quad \text{to have rank} \quad n.$$

Lemma 7.8

The DLM is observable if and only if $\left[\underline{F}^T\ \underline{R}_{n-1}^T\right]$ has rank n.

Proof

Referring back to equation (7.6), it is clear that

$$\text{rank}(\underline{R}_d) = \text{rank}\left[(\underline{FG})^T(\underline{FG}^2)^T \ldots(\underline{FG}^d)^T\right]$$

since the transformation is non-singular. Thus $\text{rank}\left[\underline{F}^T\ \underline{R}_{n-1}^T\right]$ is n if and only if the rank of the observability matrix is n.

Lemma 7.9

If the DLM is observable, then $n \leq d+1$.

Proof

If the DLM is observable, then $\text{rank}\left[\underline{F}^T\ \underline{R}_{n-1}^T\right]$ is n. The last row of $\underline{R}_{n-1}$ is $\underline{FG}(\underline{G}-\underline{I})^{n-2}$. From equation (7.3), $\underline{FG}(\underline{G}-\underline{I})^{n-2} = \underline{0}$ if $n-2 \geq d$. Hence if the model is observable, $n-2 < d$ or $n \leq d+1$.

Lemmas 7.7 and 7.9 imply that there are two possible dimensions for observable DLMs given by $n = d$ and $n = d+1$. In Section 5.4, we gave a result due to Kalman. From this we found that if the model is observable, and $\underline{W}$ is positive definite, then the system converges to an equilibrium state. Thus we see that observability is a desirable property for DLMs. Hence in the sequel, we assume that n can only take one of these two values.

Lemma 7.10

The $(d+1) \times n$ matrix $\left[\underline{F}^T, \underline{R}_d^T\right]^T$ has rank d if and only if $\underline{F}(\underline{G}-\underline{I})^d = \underline{O}$ and otherwise has rank $d+1$.

Proof

$\left[\underline{F}^T \; \underline{R}_d^T\right]$ must have rank at least d, since $\underline{R}_d$ has rank d from Lemma 7.1.

Suppose it has rank d. Then the rows of the matrix are linearly dependent, and there are constants a_i, not all zero, such that

$$a_o\underline{FG}(\underline{G}-\underline{I})^{d-1}+a_1\underline{FG}(\underline{G}-\underline{I})^{d-2}+\ldots+a_{d-1}\underline{FG}+a_d\underline{F} = \underline{O} \; . \qquad (7.23)$$

$a_d \neq 0$, since otherwise (7.23) would imply that $\underline{R}_d$ has rank less than d. Postmultiply (7.23) by $\underline{G}$, noting that $\underline{G}$ and $(\underline{G}-\underline{I})$ commute, and that

$$\underline{G}^2(\underline{G}-\underline{I})^{i-1} = \underline{G}(\underline{G}-\underline{I})^i+\underline{G}(\underline{G}-\underline{I})^{i-1} \; .$$

Then equation (7.23) becomes

$$a_o(\underline{FG}(\underline{G}-\underline{I})^d+\underline{FG}(\underline{G}-\underline{I})^{d-1})+a_1(\underline{FG}(\underline{G}-\underline{I})^{d-1}+\underline{FG}(\underline{G}-\underline{I})^{d-2})+\ldots$$

$$\ldots+a_{d-1}(\underline{FG}(\underline{G}-\underline{I})+\underline{FG})+a_d\underline{FG} = \underline{O} \; .$$

Thus

$$(a_o+a_1)\underline{FG}(\underline{G}-\underline{I})^{d-1}+(a_1+a_2)\underline{FG}(\underline{G}-\underline{I})^{d-2}+\ldots+(a_{d-1}+a_d)\underline{FG} = \underline{O} \; .$$

Since all these terms are row vectors of $\underline{R}_d$, it follows from Lemma 7.1 that they must be linearly independent, hence all the coefficients must be zero. Thus

$$a_1 = -a_o, \quad a_2 = -a_1 = a_o, \ldots \; .$$

In general

$$a_i = (-1)^i a_o \qquad i = 0,1,\ldots d \; . \qquad (7.24)$$

Since $a_d = (-1)^d a_o \neq 0$, then $a_o \neq 0$. Substituting the values of a_i given by (7.24) into equation (7.23)

$$a_o\left[\underline{F}\underline{G}(\underline{G}-\underline{I})^{d-1}-\underline{F}\underline{G}(\underline{G}-\underline{I})^{d-2}+\ldots+(-1)^{d-1}\underline{F}\underline{G}+(-1)^d\underline{F}\right] = \underline{O}$$

or $a_o\underline{F}(\underline{G}-\underline{I})^d = \underline{O}$ which implies that

$$\underline{F}(\underline{G}-\underline{I})^d = \underline{O}$$

since $a_o \neq 0$.

Conversely, if $\underline{F}(\underline{G}-\underline{I})^d = \underline{O}$, then

$$\underline{F}\underline{G}(\underline{G}-\underline{I})^{d-1}-\underline{F}\underline{G}(\underline{G}-\underline{I})^{d-2}+\ldots+(-1)^{d-1}\underline{F}\underline{G}+(-1)^d\underline{F} = \underline{O}\ .$$

Thus the rows of $\left[\underline{F}^T\ \underline{R}_d^T\right]^T$ are linearly dependent, and the matrix has rank less than $d+1$. But since it contains a matrix of rank d, it must itself have rank d.

Corollary

The condition $\underline{F}(\underline{G}-\underline{I})^d \neq \underline{O}$ implies that $n \geq d+1$.

Proof

$\underline{F}(\underline{G}-\underline{I})^d \neq \underline{O}$ means that the $(d+1) \times n$ matrix $\left[\underline{F}^T\ \underline{R}_d^T\right]^T$ has rank $d+1$, hence n must be greater than or equal to $d+1$.

Notice that $n \geq d+1$ does not imply that $\underline{F}(\underline{G}-\underline{I})^d \neq \underline{O}$. When $n = d+1$, $\left[\underline{F}^T\ \underline{R}_d^T\right]$ has the same rank as the observability matrix, and Lemma 7.8 can be rewritten as:

Corollary

When $n = d+1$, the system is observable if and only if $\underline{F}(\underline{G}-\underline{I})^d \neq \underline{O}$.

We now investigate the effect of the condition $\underline{F}(\underline{G}-\underline{I})^d = \underline{O}$ on the range of the βs.

Consider β_d, which being the product of d factors

each less than one in modulus, must itself have modulus less than one. From equation (7.18),

$$\beta_d = (-1)^d\binom{d}{d} + \sum_{i=o}^{d-1}\binom{d}{i}(-1)^i \underline{FG}^{d-i}\underline{A}$$

$$= (-1)^d - (-1)^d\underline{FA} + \sum_{i=o}^{d-1}\binom{d}{i}(-1)^i \underline{FG}^{d-i}\underline{A} + (-1)^d\underline{FA}$$

$$= (-1)^d(1-\underline{FA}) + \sum_{i=o}^{d}\binom{d}{i}(-1)^i \underline{FG}^{d-i}\underline{A}$$

$$= (-1)^d(1-\underline{FA}) + \underline{F}(\underline{G}-\underline{I})^d\underline{A}$$

so that when $\underline{F}(\underline{G}-\underline{I})^d = \underline{O}$, β_d reduces to $(-1)^d(1-\underline{FA})$. Now

$$\underline{FA} = \underline{FPF}^T/(\underline{FPF}^T+\underline{V}),$$

hence $0 < \underline{FA} < 1$. Thus we have shown the following:

Lemma 7.11

If $\underline{F}(\underline{G}-\underline{I})^d = \underline{O}$, then

if d is even, $0 < \beta_d < 1$

while if d is odd, $-1 < \beta_d < 0$.

We see that the range of values taken by β_d is effectively halved by the condition $\underline{F}(\underline{G}-\underline{I})^d = \underline{O}$. This is obviously a restriction we would wish to avoid, hence the ideal choice of model will be observable, with $\underline{F}(\underline{G}-\underline{I})^d \neq \underline{O}$ and $n = d+1$. Example 7.2 has $n = d$, and hence is restricted as in Lemma 7.11. The next result is surprising, since it eliminates several intuitive models.

Theorem 7.12

If $\underline{G}$ is non-singular, then the β parameters cannot cover the complete stability region.

Proof

From (7.3), $\underline{FG}(\underline{G}-\underline{I})^d = \underline{O}$.

If $\underline{G}$ is non-singular, then $\underline{G}^{-1}$ exists, and we can postmultiply equation (7.3) by $\underline{G}^{-1}$ to give

$$\underline{F}(\underline{G}-\underline{I})^{d} = \underline{O}$$

since $\underline{G}$ and $(\underline{G}-\underline{I})$ commute. Hence from Lemma 7.11, the parameters are restricted.

For an unrestricted model, $\underline{G}$ must be singular. In fact we have

Lemma 7.13

If $\underline{F}(\underline{G}-\underline{I})^{d} \neq \underline{O}$ and $n = d+1$, then

$$\text{rank}(\underline{G}) = n-1 = d .$$

Proof

From equations (7.8) and (7.10)

$$\begin{bmatrix} \underline{F} \\ \underline{R}_d \end{bmatrix} \underline{G} = \begin{bmatrix} \underline{FG} \\ \underline{KR}_d \end{bmatrix}$$

$$= \left[\begin{array}{c|ccc} \underline{O} & 1 & 0 & \cdots\cdots\ 0 \\ \hline \underline{O} & & \underline{K} & \end{array}\right] \begin{bmatrix} \underline{F} \\ \underline{R}_d \end{bmatrix} = \underline{P} \begin{bmatrix} \underline{F} \\ \underline{R}_d \end{bmatrix}$$

where $\underline{K}$ is the matrix defined after equation (7.10). Since $\left[\underline{F}^{T}\ \underline{R}_d^{T}\right]^{T}$ is non-singular by Lemma 7.10,

$$\text{rank}(\underline{G}) = \text{rank}(\underline{P}) = n-1 = d$$

Corollary

If $n = d$, $\text{rank}(\underline{G}) = d$.

Proof

$\underline{R}_d\underline{G} = \underline{KR}_d$.

Since $\underline{R}_d$ is non-singular, $\text{rank}(\underline{G}) = \text{rank}(\underline{K}) = n$.

Example 7.3

An example of a polynomial model of degree d-1 is given by

$$\underline{F} = \underline{1}_{d+1}^T, \quad \underline{G} = \begin{bmatrix} \underline{1}_{d+1} & \underline{1}_d & \cdots & \underline{1}_2 & \underline{0} \end{bmatrix}$$

where

$$\underline{1}_r = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{matrix} \} \, d+1-r \\ \\ \\ \\ \} \, r \end{matrix}$$

Thus $\underline{G}$ can be written $\begin{bmatrix} \underline{\Omega} & \underline{0} \\ \underline{1}_d^T & 0 \end{bmatrix}$

where

$$\underline{\Omega} = \begin{bmatrix} 1 & 0 & . & . & . & 0 \\ 1 & 1 & . & . & . & 0 \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ 1 & 1 & . & . & . & 1 \end{bmatrix}.$$

$\underline{\Omega}$ satisfies $(\underline{\Omega}-\underline{I})^{d-1} \neq \underline{0}$, $(\underline{\Omega}-\underline{I})^d = \underline{0}$ and $\underline{\Omega}(\underline{\Omega}-\underline{I})^{d-1} \neq \underline{0}$. Notice that $\underline{\Omega}$ is the transpose of the polynomial model proposed by Harrison and Stevens (1976). It is easily shown that

$$\underline{G}(\underline{G}-\underline{I})^{d-1} = \begin{bmatrix} \underline{\Omega}(\underline{\Omega}-\underline{I})^{d-1} & \underline{0} \\ \underline{1}_d^T(\underline{\Omega}-\underline{I})^{d-1} & 0 \end{bmatrix}$$

so that $\underline{F}\underline{G}(\underline{G}-\underline{I})^{d-1} \neq \underline{0}$.

It can also be shown that

$$(\underline{G}-\underline{I})^d = \begin{bmatrix} \underline{0} & \underline{0} \\ (-1)^{d+1}\underline{1}_d^T\underline{\Omega}^{-1} & (-1)^d \end{bmatrix}.$$

Hence $\underline{F}(\underline{G}-\underline{I})^d \neq \underline{0}$ and $\underline{G}(\underline{G}-\underline{I})^d = \underline{0}$.

Thus this model is a polynomial model of degree d-1, and since $\underline{F}(\underline{G}-\underline{I})^d \neq \underline{0}$ and n = d+1, the model is observable. Hence if $\underline{W}_t$ is positive definite, the model converges to an equilibrium solution, and the theory of Section 7.2 is applicable.

Example 7.4

Putting $d = 2$ in Example 7.3 yields a polynomial model of degree one, with $\underline{G}$ of dimension 3. We have

$$y_t = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \underline{\theta}_t + \varepsilon_t$$

$$\underline{\theta}_t = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix} \underline{\theta}_{t-1} + \underline{w}_t$$

where

$$\underline{W}_t \quad \begin{bmatrix} w_1 & & \\ & w_2 & \\ & & w_3 \end{bmatrix}$$

is diagonal and $\text{var}(\varepsilon_t) = v$. Hence

$$\underline{FG} = \begin{bmatrix} 3 & 2 & 0 \end{bmatrix}$$

$$\underline{FG}(\underline{G}-\underline{I}) = \begin{bmatrix} 2 & 0 & 0 \end{bmatrix}$$

$$\underline{F}(\underline{G}-\underline{I})^2 = \begin{bmatrix} 0 & -1 & 1 \end{bmatrix} .$$

From Example 7.3, there is an equilibrium solution. Under steady state conditions, the Kalman updating equation is

$$\hat{\underline{\theta}}_t = \underline{G}\hat{\underline{\theta}}_{t-1} + \underline{A}(y_t - \underline{FG}\hat{\underline{\theta}}_{t-1})$$

where

$$\underline{A} = \begin{bmatrix} A_1 & A_2 & A_3 \end{bmatrix}^T .$$

From equation (7.17)

$$\beta_1 = -2+\underline{FGA} = -2+3A_1+2A_2$$

$$\beta_2 = 1-\underline{FGA}+\underline{FG}(\underline{G}-\underline{I})\underline{A} \quad = 1-3A_1-2A_2+2A_1 = 1-A_1-2A_2 .$$

Referring back to Example 5.1, the stability conditions for this model are

(i) $1+\beta_1+\beta_2 > 0$

(ii) $1-\beta_1+\beta_2 > 0$

(iii) $\beta_2 < 1$.

The steady state equation for the covariance matrix C_t is difficult to solve analytically for this model, so we shall attempt to find conditions under which (i), (ii) and (iii) are satisfied for all time t.

For convenience, we write

$$\underline{C}_{t-1} = \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix} .$$

Using the Kalman updating equations (4.4) - (4.8), we find that $\underline{A}_t = \begin{bmatrix} A_{1t} & A_{2t} & A_{3t} \end{bmatrix}^T$ is given by

$$A_{1t} = (3a+2b+w_1)/D_t$$

$$A_{2t} = (3a+5b+2d+w_2)/D_t$$

$$A_{3t} = (3a+5b+2d+w_3)/D_t$$

where

$$D_t = 9a+12b+4d+w_1+w_2+w_3+v .$$

Thus in the time dependent case,

(iii) $$\beta_2 = 1-(A_{1t}+2A_{2t})$$

$$= \frac{-w_2+w_3+v}{9a+12b+4d+w_1+w_2+w_3+v}$$

which clearly lies between -1 and 1.

(i) $$1+\beta_1+\beta_2 = 2A_1$$

$$= \frac{2(3a+2b+w_1)}{9a+12b+4d+w_1+w_2+w_3+v}$$

(ii) $$1-\beta_1+\beta_2 = \frac{4(3a+5b+2d+w_3+v)}{9a+12b+4d+w_1+w_2+w_3+v} \quad .$$

It is not clear whether constraints (i) and (ii) are satisfied.

Using equation (4.8), we can find expressions for the elements of $\underline{C}_t$ in terms of those of $\underline{C}_{t-1}$. In the following, $(\text{expression})_t$ is taken to mean the expression at time t. Where the subscript (t-1) appears on the right, it is suppressed. From these values, we find that

$$D_t(3a+2b+w_1)_t =$$

$$8(ad-b^2)+8w_1(a+2b+d)+w_1(9a+12b+4d+w_1+2w_2+4(w_3+v))$$

$$+4a(w_3+v)+(a+2b)(w_3+v-w_2) \quad . \qquad (7.25)$$

All these terms are positive, except possibly the last one, which we examine in more detail.

$$D_{t-1}(a+2b)_{t-1} = (3a+2b+w_1)_{t-2}(w_3+v-w_2)$$

so that if $(3a+2b+w_1)_{t-2}$ is greater than zero, then $(a+2b)_{t-1}$ has the same sign as (w_3+v-w_2) and the last term of (7.25) is positive.

Thus sufficient conditions for $1+\beta_1+\beta_2 > 0$ for all t is given by

$$(3a+2b+w_1)_o > 0$$

$$(3a+2b+w_1)_1 > 0 \quad .$$

Similarly

$$D_t(3a+5b+2d+w_3+v)_t =$$

$$4(ad-b^2)+4w_1(a+2b+d)+4w_1(w_3+v)$$

$$+(w_3+v)(19a+21b+6d+w_3+v)+w_2(a+3b+2d+3(w_3+v)).$$

We see that the first four terms are positive, and by the same method, the fifth term is positive if $(3a+5b+2d+w_3+v)_{t-2}$ is positive. Thus the Example 7.4 has its parameters inside the triangular stability region for all t if the following initial conditions apply

$$(3a+2b+w_1)_o > 0$$

$$(3a+2b+w_1)_1 > 0$$

$$(3a+5b+2d+w_3+v)_o > 0$$

$$(3a+5b+2d+w_3+v)_1 > 0 \quad .$$

The only term which could be negative in any of these inequalities is the covariance b, since all the other variables are variances and hence positive. Hence it seems that these conditions are not very restrictive. Indeed, in many cases, the initial estimate of b is zero. Note that the conditions for stability given above are sufficient, and not necessary.

As before, we can write the model in the form of a Box-Jenkins model,

$$\nabla^2 y_t = w_{1t} + w_{1t-1} + w_{2t} - w_{2t-2} + \nabla^2(w_{3t}+\varepsilon_t)$$

which has the same correlation structure as

$$\nabla^2 y_t = a_t + \beta_1 a_{t-1} + \beta_2 a_{t-2} \quad .$$

Multiply $\nabla^2 y_t$ in turn by $\nabla^2 y_t$, $\nabla^2 y_{t-1}$, $\nabla^2 y_{t-2}$, and take expectations, then equating the results of the two equations

$$E\left[(\nabla^2 y_t)^2\right] = 2w_1+2w_2+6(w_3+v) = (1+\beta_1^2+\beta_2^2)\,var(a)$$

$$E\left[(\nabla^2 y_t)(\nabla^2 y_{t-1})\right] = w_1-4(w_3+v) = \beta_1(1+\beta_2)\,var(a)$$

$$E\left[(\nabla^2 y_t)(\nabla^2 y_{t-2})\right] = -w_2+w_3+v = \beta_2 var(a) .$$

These have solution

$$w_1 = var(a)(1+\beta_1+\beta_2)^2/4$$

$$w_2 = var(a)(-\beta_2+(1-\beta_1+\beta_2)^2/16)$$

$$w_3+v = var(a)(1-\beta_1+\beta_2)^2/16$$

from which

$$(1-\beta_1+\beta_2)^2 > 16\beta_2 .$$

This appears to be a somewhat larger region than that covered by the Harrison-Stevens linear growth model, but in the steady state it still does not cover the complete region.

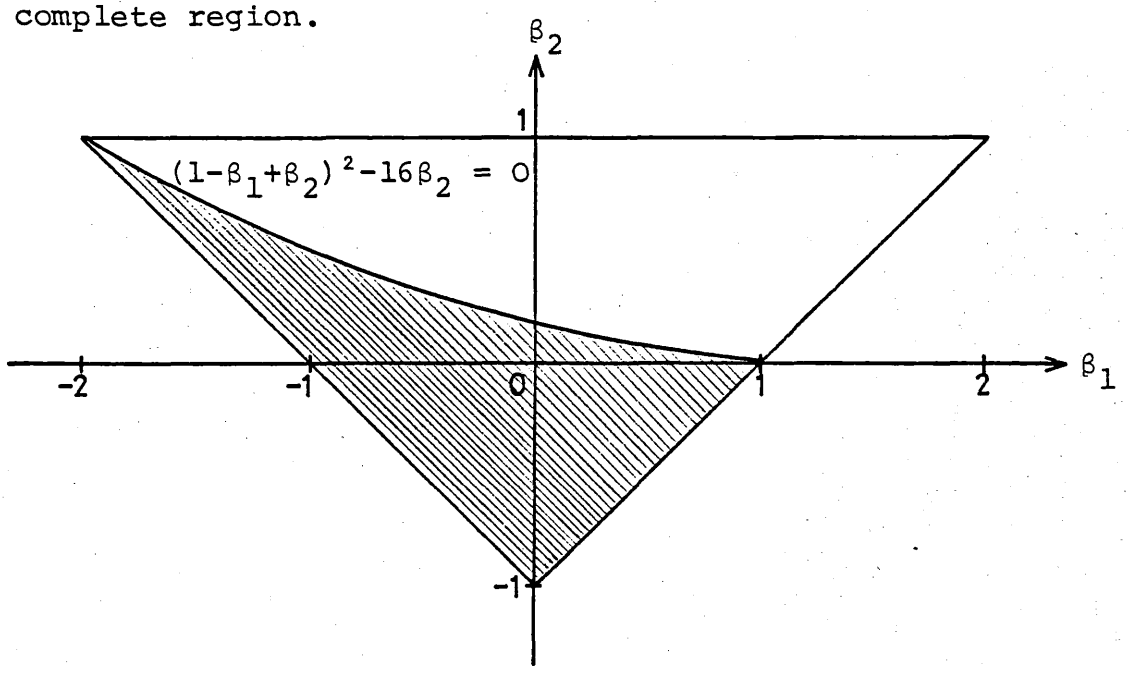


FIGURE 7.3: STABILITY REGION FOR EXAMPLE 7.4

CHAPTER 8

FORWARD-SHIFTED POLYNOMIAL PREDICTOR MODELS.

8.1 Definition of Model.

We now extend the ideas of the previous chapter to include DLMs for which the forecast function $\{y_t(k);k=1,2,...\}$ does not describe a polynomial path for short lead times, but for all lead times greater than some positive integer r, the forecast function is a polynomial of degree $d-1$. We assume that r is the smallest such integer. Thus for $k > r > 0$, the k-step ahead forecast is given by

$$y_t(k) = a_o + a_1 k + \ldots + a_{d-1} k^{d-1} \qquad (8.1)$$

where $a_{d-1} \neq 0$, while for $1 \leq k \leq r$, there is no representation for $y_t(k)$ of the form $\sum_{i=o}^{m} a_i k^i$ where m is any positive integer and $a_m \neq 0$.

Thus $y_t(k)$ is given by a curve of degree $d-1$ passing through the points $y_t(r+1),\ldots,y_t(r+d)$. All the polynomial models of Chapter 7 correspond to this description with $r = 0$, so that in these cases the polynomial curve passes through the points $y_t(1),\ldots,y_t(d),\ldots$. In the models of this chapter, however, the polynomial is shifted by an amount $r > 0$ and $y_t(1),\ldots,y_t(r)$ are not on the curve. For example, consider a linear forecast model ($d = 2$) with shift $r = 1$. Then the curve is a straight line, as in Figure 8.1.

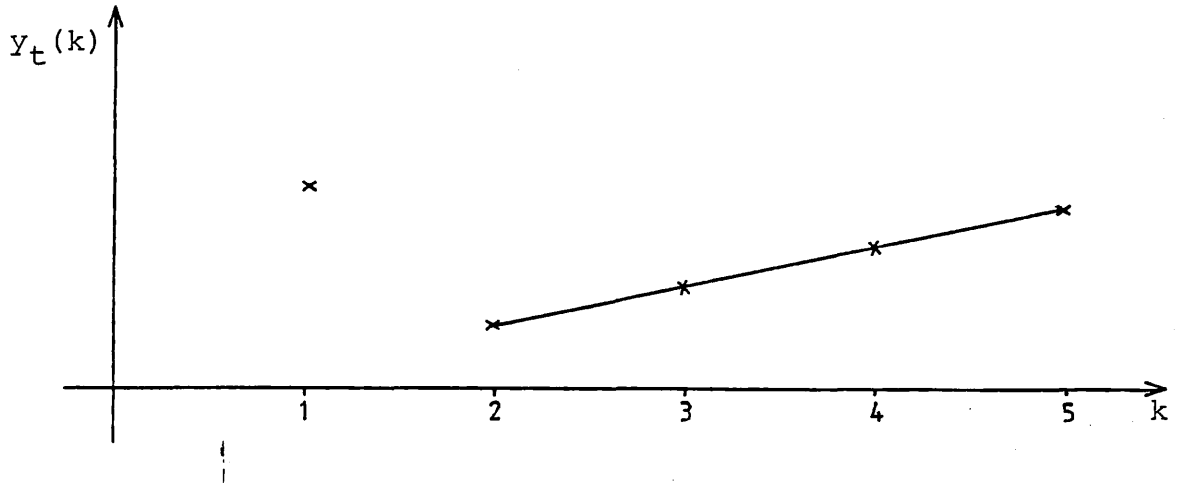


FIGURE 8.1: EXAMPLE OF AN EVENTUAL FORECAST FUNCTION FOR AN ARIMA (0,2,3) PROCESS.

This line passes through the points $y_t(2), y_t(3), \ldots$ but $y_t(1)$ is not on the line. Thus there is a 'jump' from $y_t(1)$ to $y_t(2)$ which cannot be expressed in terms of the linear formulation

$$y_t(k) = a_o + a_1 k$$

which holds for $k \geq 2$.

As in Chapter 7, the representation (8.1) is equivalent to

$$\nabla^d y_t(k) \equiv 0$$

or

$$\sum_{i=o}^{d} \binom{d}{i} (-1)^i y_t(k-i) \equiv 0 \quad k \geq d + r + 1. \tag{8.2}$$

Since the predictor of the DLM (4.17) is given by

$$y_t(k) = \underline{F} G^k \hat{\underline{\theta}}_t,$$

equation (8.2) becomes

$$\sum_{i=o}^{d} \binom{d}{i} (-1)^i \underline{F} G^{k-i} \hat{\underline{\theta}}_t \equiv 0 \qquad k \geq d + r + 1 .$$

This holds for all values of $\hat{\underline{\theta}}_t$, so we have

$$\underline{F}\underline{G}^{r+j}(\underline{G}-\underline{I})^{d} = \underline{O} \qquad j \geq 1 .$$

In particular

$$\underline{F}\underline{G}^{r+1}(\underline{G}-\underline{I})^{d} = \underline{O}. \tag{8.3}$$

As before, $a_{d-1} \neq 0$ implies that

$$\underline{F}\underline{G}^{r+1}(\underline{G}-\underline{I})^{d-1} \neq \underline{O}. \tag{8.4}$$

Since r is the smallest integer for which (8.3) holds,

$$\underline{F}\underline{G}^{r}(\underline{G}-\underline{I})^{d} \neq \underline{O} . \tag{8.5}$$

Equation (8.3) is clearly an extension of equation (7.3). Notice that $\underline{G}$ must be singular, otherwise (8.5) cannot hold. In order to find $y_t(k)$ for all $k \geq 1$, it is necessary only to find $y_t(1), y_t(2), \ldots y_t(d+r)$. $y_t(k)$, $k \geq d + r + 1$ can then be found recursively from equation (8.2). It is convenient to consider instead an invertible linear transformation of these quantities, namely

$$\underline{F}\underline{G}\hat{\theta}_t, \underline{F}\underline{G}^{2}\hat{\theta}_t, \ldots, \underline{F}\underline{G}^{r+1}\hat{\theta}_t, \underline{F}\underline{G}^{r+1}(\underline{G}-\underline{I})\hat{\theta}_t, \ldots, \underline{F}\underline{G}^{r+1}(\underline{G}-\underline{I})^{d-1}\hat{\theta}_t .$$

Define

$$\underline{R}_{r,d} = \left[(\underline{F}\underline{G})^{T}(\underline{F}\underline{G}^{2})^{T}\ldots(\underline{F}\underline{G}^{r+1})^{T}(\underline{F}\underline{G}^{r+1}(\underline{G}-\underline{I}))^{T}\ldots(\underline{F}\underline{G}^{r+1}(\underline{G}-\underline{I})^{d-1})^{T}\right]^{T} \tag{8.6}$$

Then $\underline{R}_{r,d}\hat{\underline{\theta}}_t$ is an invertible transformation of $y_t(1), \ldots, y_t(d+r)$. None of the rows of $\underline{R}_{r,d}$ can be zero, because this would imply that $y_t(k)$ is a polynomial of degree less than $d-1$.

Lemma 8.1

The rank of the $(r+d) \times n$ matrix $\underline{R}_{r,d}$ is $r+d$.

Proof.

This is proved in the same manner as Lemma 7.1. The equation to be postmultiplied is slightly more complicated,

and there are two distinct cases, $k < r$ and $k \geq r+1$, where k is defined in the proof of Lemma 7.1.

We now premultiply the Kalman updating equation (4.4) by $\underline{R}_{r,d}$, to obtain

$$\underline{R}_{r,d}\hat{\underline{\theta}}_t = \underline{R}_{r,d}\underline{G}\hat{\underline{\theta}}_{t-1} + \underline{R}_{r,d}\underline{A}_t(y_t - \underline{F}\underline{G}\hat{\underline{\theta}}_{t-1}). \tag{8.7}$$

Furthermore, using

$$\underline{F}\underline{G}^{r+2}(\underline{G}-\underline{I})^i = \underline{F}\underline{G}^{r+1}(\underline{G}-\underline{I})^{i+1} + \underline{F}\underline{G}^{r+1}(\underline{G}-\underline{I})^i,$$

$\underline{R}_{r,d}\underline{G}$ can be expressed as

$$\begin{bmatrix} \underline{F}\underline{G}^2 \\ \underline{F}\underline{G}^3 \\ \cdot \\ \cdot \\ \cdot \\ \underline{F}\underline{G}^{r+1} \\ \underline{F}\underline{G}^{r+2} \\ \underline{F}\underline{G}^{r+2}(\underline{G}-\underline{I}) \\ \cdot \\ \cdot \\ \cdot \\ \underline{F}\underline{G}^{r+2}(\underline{G}-\underline{I})^{d-1} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdot & \cdot & \cdot & & 0 \\ & \cdot & & & & & & \\ & & \cdot & & & & & \\ & & & \cdot & & & & \\ & & & & \cdot & 1 & & \\ & & & & & 0 & 1 & \\ & & & & & & 1 & 1 \\ & & & & & & & \cdot \\ & & & & & & & & \cdot \\ & & & & & & & & & \cdot \\ & & & & & & & & & & 1 \\ 0 & & & \cdot & \cdot & \cdot & & 0 & 1 \end{bmatrix} \begin{bmatrix} \underline{F}\underline{G} \\ \underline{F}\underline{G}^2 \\ \cdot \\ \cdot \\ \cdot \\ \underline{F}\underline{G}^{r} \\ \underline{F}\underline{G}^{r+1} \\ \underline{F}\underline{G}^{r+1}(\underline{G}-\underline{I}) \\ \cdot \\ \cdot \\ \cdot \\ \underline{F}\underline{G}^{r+1}(\underline{G}-\underline{I})^{d-1} \end{bmatrix}.$$

Writing

$$m_{it} = \underline{F}\underline{G}^i\hat{\underline{\theta}}_t, \quad \mu_{it} = \underline{F}\underline{G}^i\underline{A}_t \qquad i = 1,\ldots,r$$

$$m_{it} = \underline{F}\underline{G}^{r+1}(\underline{G}-\underline{I})^{i-r-1}\hat{\underline{\theta}}_t, \mu_{it} = \underline{F}\underline{G}^{r+1}(\underline{G}-\underline{I})^{i-r-1}\underline{A}_t \quad i=r+1,\ldots,r+d$$

and

$$\underline{M}_t = (m_{1t},\ldots,m_{d+r,t})^T, \quad \underline{\mu}_t = (\mu_{1t},\ldots,\mu_{d+r,t})^T$$

we find that (8.7) becomes

$$\underline{M}_t = \underline{K}\underline{M}_{t-1} + \underline{\mu}_t(y_t - m_{1t-1}) \tag{8.8}$$

where the elements of $\underline{K} = (K_{ij})$ are given by

$$K_{ij} = \begin{cases} 1 & i = j-1 \qquad j=2,\ldots d+r \\ 1 & i = j \qquad\quad j \geq r+1 \\ 0 & \text{elsewhere} \end{cases}$$

Comparing (8.8) with the work of Godolphin and Harrison (1973), we see that the first r rows and the last d rows of this equation have the same formulation as equations (3.10) and (3.12) respectively of the above paper. We now consider equation (3.11) of Godolphin and Harrison (1973) in the terminology of this chapter. Thus we derive

Lemma 8.2

$$y_t(k) = \underline{FG}^k \hat{\underline{\theta}}_t = \sum_{j=1}^{d} \binom{k-r-1}{j-1} m_{jt} \qquad k \geq r+1 .$$

Proof

$$\sum_{j=1}^{d} \binom{k-r-1}{j-1} m_{jt} = \sum_{j=1}^{d} \binom{k-r-1}{j-1} \underline{FG}^{r+1} (\underline{G}-\underline{I})^{j-1} \hat{\underline{\theta}}_t \qquad k \geq r+1$$

which is the same expression as that which occurs in Lemma 7.2, with k replaced by $k-r$, and $\underline{FG}$ replaced by $\underline{FG}^{r+1}$.

For $k - r < d$, the summation reduces to

$$\sum_{j=1}^{k-r} \binom{k-r-1}{j-1} m_{jt} .$$

Following the argument of Lemma 7.2 we have for

$r+1 \leq k \leq r+d$

$$\sum_{j=1}^{k-r} \binom{k-r-1}{j-1} m_{jt} = \sum_{i=1}^{k-r} \underline{FG}^{r+i} \binom{k-r-1}{i-1} \sum_{j=o}^{k-r-i} \binom{k-r-i}{k-r-i-j} (-1)^j \hat{\underline{\theta}}_t$$

$$= \underline{FG}^{r+k-r} \hat{\underline{\theta}}_t$$

$$= y_t(k) .$$

The lemma follows for all $k \geq r + 1$ from equation (8.2).

These results suggest that we are dealing with an ARIMA (0,d,d+r) process, $r > 0$, but before we can state this, we need to check the range of values that can be taken by $\underline{\mu}_t$ in this model.

8.2 Steady State Theory

As before, since we wish to compare the DLM with the static ARIMA model, we assume steady state conditions, so that equation (8.8) becomes

$$\underline{M}_t = \underline{K}\underline{M}_{t-1} + \underline{\mu}(y_t - m_{1t-1}) \quad . \qquad (8.9)$$

Taking z-transforms of equation (8.9), we have

$$\underline{M}(z) = z^{-1}\underline{K}\underline{M}(z) + \underline{\mu}(Y(z) - z^{-1}M_1(z)). \qquad (8.10)$$

We need to solve this set of $d + r$ equations to find $M_i(z)/Y(z)$ in terms of the elements of $\underline{\mu}$.

Lemma 8.3

Equations (8.10) satisfy

$$M_k(z) = (zY(z) - M_1(z))\left[\sum_{i=1}^{d} \mu_{r+i}(z-1)^{-i}z^{k-1-r} + \sum_{i=k}^{r} \mu_i z^{k-1-i}\right]$$

$$1 \leq k \leq r \qquad (8.11)$$

$$M_k(z) = (zY(z) - M_1(z))\left[\sum_{i=k}^{d+r} \mu_i (z-1)^{k-1-i}\right] \quad r+1 \leq k \leq d+r \qquad (8.12)$$

Proof

From equation (8.10), we have

$$M_i(z) = z^{-1}M_{i+1}(z) + \mu_i(Y(z) - z^{-1}M_1(z)) \qquad 1 \leq i \leq r$$

$$M_i(z) = z^{-1}(M_i(z) + M_{i+1}(z)) + \mu_i(Y(z) - z^{-1}M_1(z))$$

$$r + 1 \leq i \leq r + d - 1$$

$$M_{d+r}(z) = z^{-1}M_{d+r}(z) + \mu_{d+r}(Y(z) - z^{-1}M_1(z)).$$

From the last of these equations (8.12) holds for $k = d + r$. We can then prove by induction as in Lemma 7.3 that (8.12) holds for $r + 1 \leq k \leq d + r$, and in particular

$$M_{r+1}(z) = (zY(z) - M_1(z)) \sum_{i=r+1}^{d+r} \mu_i(z-1)^{r-i} .$$

Substituting this value into the above expression for $M_r(z)$, we have

$$zM_r(z) = (zY(z) - M_1(z)) \sum_{i=r+1}^{d+r} \mu_i(z-1)^{r-i} + \mu_r(zY(z) - M_1(z))$$

$$M_r(z) = (zY(z) - M_1(z)) \sum_{i=r}^{d+r} \mu_i(z-1)^{r-i}z^{-1}$$

which is (8.11) with $k = r$. We proceed by induction to prove (8.11) for $1 \leq k \leq r$, and the result follows.

Lemma 8.4.

The characteristic polynomial of the model, is given by

$$\beta(z) = \sum_{i=o}^{d+r} \beta_i z^{d+r-i}$$

$$= z^r(z-1)^d + \sum_{i=1}^{r} \mu_i z^{r-i}(z-1)^d + \sum_{i=1}^{d} \mu_{r+i}(z-1)^{d-i}. \qquad (8.13)$$

Proof

The characteristic polynomial of the model is the denominator of $M_i(z)/Y(z)$. Putting $k = 1$ in equation (8.11)

$$M_1(z) = (zY(z)-M_1(z))\left[\sum_{i=1}^{d} \mu_{r+i}(z-1)^{-i}z^{-r} + \sum_{i=1}^{r} \mu_i z^{-i}\right].$$

Collecting terms

$$M_1(z)\ 1 + \left[\sum_{i=1}^{d} \mu_{r+i}(z-1)^{-i}z^{-r} + \sum_{i=1}^{r} \mu_i z^{-i}\right]$$

$$= zY(z)\left[\sum_{i=1}^{d} \mu_{r+i}(z-1)^{-i}z^{-r} + \sum_{i=1}^{r} \mu_i z^{-i}\right],$$

or

$$\frac{M_1(z)}{Y(z)} = \frac{z^{r+1}(z-1)^d\left[\sum_{i=1}^{d} \mu_{r+i}(z-1)^{-i}z^{-r} + \sum_{i=1}^{r} \mu_i z^{-i}\right]}{z^r(z-1)^d\left[1+\sum_{i=1}^{d} \mu_{r+i}(z-1)^{-i}z^{-r} + \sum_{i=1}^{r} \mu_i z^{-i}\right]}.$$

Thus the characteristic polynomial of the model is

$$\beta(z) = z^r(z-1)^d + \sum_{i=1}^{r} \mu_i z^{r-i}(z-1)^d + \sum_{i=1}^{d} \mu_{r+i}(z-1)^{d-i}$$

as required.

Lemma 8.5

β_i is given by

$$\beta_i = \binom{d}{i}(-1)^i + \sum_{j=o}^{i-1} \binom{d}{j}(-1)^j \underline{F}\underline{G}^{i-j}\underline{A} \qquad i=1,\ldots d+r \tag{8.14}$$

where $\beta_o = 1$, and we adhere to the convention that $\binom{a}{b} = o$ for $b > a$.

Proof

We need to express (8.13) as $\sum_{i=o}^{d+r} \beta_i z^{d+r-i}$. Taking (8.13) term by term

$$z^r(z-1)^d = \sum_{j=o}^{d} \binom{d}{j}(-1)^j z^{d+r-j} \tag{8.15}$$

$$\sum_{i=1}^{r} \mu_i z^{r-i}(z-1)^d = \sum_{i=1}^{r} \sum_{j=o}^{d} \mu_i \binom{d}{j} (-1)^j z^{d+r-i-j}$$

$$= \sum_{i=1}^{r} \sum_{k=i}^{d+i} \mu_i \binom{d}{k-i} (-1)^{k-i} z^{d+r-k}$$

$$= \sum_{k=1}^{d+r} z^{d+r-k} \sum_{i=\max(1,k-d)}^{\min(r,k)} \mu_i \binom{d}{k-i} (-1)^{k-i} . \quad (8.16)$$

Assume $r < d$. Then the right hand side of (8.16) can be written

$$\sum_{k=1}^{r} z^{d+r-k} \sum_{i=1}^{k} \mu_i \binom{d}{k-i} (-1)^{k-i} + \sum_{k=r+1}^{d} z^{d+r-k} \sum_{i=1}^{r} \mu_i \binom{d}{k-i} (-1)^{k-i}$$

$$+ \sum_{k=d+1}^{d+r} z^{d+r-k} \sum_{i=k-d}^{r} \mu_i \binom{d}{k-i} (-1)^{k-i} . \quad (8.17)$$

The final term of equation (8.13) is

$$\sum_{i=1}^{d} \mu_{r+i}(z-1)^{d-i} = \sum_{i=1}^{d} \mu_{r+i} \sum_{j=o}^{d-i} \binom{d-i}{j} (-1)^j z^{d-i-j}$$

$$= \sum_{k=r+1}^{d+r} z^{d+r-k} \sum_{i=1}^{k-r} \mu_{r+i} \binom{d-i}{k-i-r} (-1)^{k-i-r} . \quad (8.18)$$

Substituting the expressions (8.15),(8.17) and (8.18) into (8.13), then for $r < d$,

$$\beta(z) = z^{d+r} + \sum_{k=1}^{r} z^{d+r-k} \left[\binom{d}{k} (-1)^k + \sum_{i=1}^{k} \mu_i \binom{d}{k-i} (-1)^{k-i} \right]$$

$$+ \sum_{k=r+1}^{d} z^{d+r-k} \left[\binom{d}{k} (-1)^k + \sum_{i=1}^{r} \mu_i \binom{d}{k-i} (-1)^{k-i} + \right.$$

$$\left. \sum_{i=1}^{k-r} \mu_{r+i} \binom{d-i}{k-i-r} (-1)^{k-i-r} \right]$$

$$+ \sum_{k=d+1}^{d+r} z^{d+r-k} \left[\sum_{i=k-d}^{r} \mu_i \binom{d}{k-i} (-1)^{k-i} + \sum_{i=1}^{k-r} \mu_{r+i} \binom{d-i}{k-i-r} (-1)^{k-i-r} \right].$$

Consequently

$$\beta(z) = \sum_{i=o}^{d+r} \beta_i z^{d+r-i}$$

where $\beta_o = 1$ and

$$\beta_i = \binom{d}{i}(-1)^i + \sum_{j=1}^{i} \underline{FG}^j \underline{A} \binom{d}{i-j} (-1)^{i-j} \qquad i = 1,\ldots,r$$

$$\beta_i = \binom{d}{i}(-1)^i + \sum_{j=1}^{r} \underline{FG}^j \underline{A} \binom{d}{i-j} (-1)^{i-j}$$

$$+ \sum_{j=1}^{i-r} \underline{FG}^{r+1} (\underline{G}-\underline{I})^{j-1} \underline{A} \binom{d-j}{i-j-r} (-1)^{i-j-r} \qquad i = r+1,\ldots,d$$

$$\beta_i = \sum_{j=i-d}^{r} \underline{FG}^j \underline{A} \binom{d}{i-j} (-1)^{i-j} + \sum_{j=1}^{i-r} \underline{FG}^{r+1} (\underline{G}-\underline{I})^{j-1} \underline{A} \binom{d-j}{i-j-r} (-1)^{i-j-r}$$

$$i = d+1,\ldots,d+r.$$

Using a similar argument to that used in the proof of Lemma 7.5,

$$\underline{FG}^{r+1} \sum_{j=1}^{i-r} (\underline{G}-\underline{I})^{j-1} \underline{A} \binom{d-j}{i-j-r} (-1)^{i-j-r} = \sum_{j=o}^{i-r-1} \binom{d}{j} (-1)^j \underline{FG}^{i-j} \underline{A} .$$

Also

$$\sum_{j=1}^{r} \underline{FG}^j \underline{A} \binom{d}{i-j} (-1)^{i-j} = \sum_{j=i-r}^{i-1} \binom{d}{j} (-1)^j \underline{FG}^{i-j} \underline{A} .$$

Hence β_i is given by (8.14) for all i, as stated in the lemma.

If we follow the same arguments for $r = d$ and $r > d$, we arrive at the same result. This definition of β_i is exactly the same form as for the non-forward shifted polynomial case (although if the same number of βs are to be found, then of course their values will be different, since d is different).

We are now in a position to prove

Theorem 8.6

Suppose the DLM (4.17) satisfies $\underline{FG}^{r+1}(\underline{G}-\underline{I})^{d} = \underline{O}$, $\underline{FG}^{r}(\underline{G}-\underline{I})^{d} \neq \underline{O}$, $\underline{FG}^{r+1}(\underline{G}-\underline{I})^{d-1} \neq \underline{O}$. Then in the steady state situation, the predictor $y_t(k) = \underline{FG}^{k}\hat{\underline{\theta}}_t$ is identical to the k-step ahead predictor of an ARIMA (0,d,d+r) process $(r > 0)$ if and only if the model's estimation scheme is stable in the steady state.

Proof

The proof of this theorem follows directly from Theorem 3.3. Setting $p = 0$ in this theorem yields

$$\phi_j = \binom{d}{j}(-1)^j$$

so that equation (3.33) is identical to equation (8.2). Also, the values of $\beta_1,\ldots,\beta_{d+r}$ in that theorem are given by

$$\beta_j = \sum_{i=0}^{j} \binom{d}{i}(-1)^i \lambda_{j-i} \qquad 1 \leq j \leq d$$

$$\beta_{j+d} = \sum_{i=0}^{d} \binom{d}{i}(-1)^i \lambda_{d+j-i} \qquad 1 \leq j \leq r$$

which, with $\lambda_j = \underline{FG}^{j}\underline{A}$ $(1 \leq j \leq d+r)$ and $\lambda_0 = 1$, are identical to equation (8.14). Substituting this value for λ_j into equation (3.33), we have, in DLM notation

$$\underline{FG}^{k}\hat{\underline{\theta}}_t = \underline{FG}^{k+1}\hat{\underline{\theta}}_{t-1} + \underline{FG}^{k}\underline{A}(y_t - \underline{FG}\hat{\underline{\theta}}_{t-1})$$

which is the Kalman updating equation (4.4) premultiplied by $\underline{FG}^{k}$, and hence holds for all $k \geq 1$. Thus the stability condition for the estimation scheme for the DLM, that the roots of $\sum_{i=0}^{d+r} \beta_i z^{d+r-i}$ should lie inside the unit circle,

are equivalent to the invertibility condition of the ARIMA (0,d,d+r) process, that all zeros of $\sum_{i=0}^{d+r} \beta_i z^i$ should lie outside the unit circle.

8.3 Size and Structure of G

From Lemma 8.1, the dimension of $\underline{G}$ must be greater than or equal to $d + r$. Thus if the first r forecasts do not follow a polynomial path, then there must be r more system variables than we would expect from the degree of the polynomial.

Lemma 8.7

The $(d+r+1) \times n$ matrix $\left[\underline{F}^T \; \underline{R}_{r,d}^T\right]^T$ has rank $d+r+1$.

Proof

We know that $\left[\underline{F}^T \; \underline{R}_{r,d}^T\right]^T$ must have rank at least $d+r$, since it contains $\underline{R}_{r,d}$ which has rank $d+r$. Suppose the matrix has rank $d+r$. Then the rows of the matrix are linearly dependent, and there are constants a_i, not all zero, such that

$$\sum_{i=o}^{r} a_i \underline{FG}^i + \sum_{i=1}^{d} a_{r+i} \underline{FG}^{r+1} (\underline{G}-\underline{I})^{i-1} = \underline{O} \; . \qquad (8.19)$$

Notice that $a_o \neq 0$, otherwise (8.19) implies that the rows of $\underline{R}_{r,d}$ are linearly dependent, which contradicts Lemma 8.1. Postmultiplying equation (8.19) by $\underline{G}$, we obtain

$$\sum_{i=o}^{r} a_i \underline{FG}^{i+1} + \sum_{i=1}^{d} a_{r+i} \left[\underline{FG}^{r+1} (\underline{G}-\underline{I})^i + \underline{FG}^{r+1} (\underline{G}-\underline{I})^{i-1}\right] = \underline{O}$$

from which

$$\sum_{i=1}^{r+1} a_{i-1} \underline{FG}^i + \sum_{i=1}^{d-1} a_{r+i} \underline{FG}^{r+1} (\underline{G}-\underline{I})^i + \sum_{i=0}^{d-1} a_{r+i+1} \underline{FG}^{r+1} (\underline{G}-\underline{I})^i = \underline{O}$$

so that

$$\sum_{i=1}^{r} a_{i-1}\underline{FG}^{i} + \sum_{i=o}^{d-1} (a_{r+i} + a_{r+i+1})\underline{FG}^{r+1}(\underline{G}-\underline{I})^{i} = \underline{O}.$$

But all these terms are row vectors of $\underline{R}_{r,d}$ and it follows from Lemma 8.1, that they are linearly independent, hence all the coefficients must be zero. In particular, $a_o = 0$, so that (8.19) cannot be true, and $\left[\underline{F}^T \ \underline{R}_{r,d}^T\right]^T$ must have rank d+r+1.

Corollary

The dimension of $\underline{G}$ is not less than d+r+1.

Proof

The (d+r+1) x n matrix $\left[\underline{F}^T \ \underline{R}_{r,d}^T\right]^T$ has rank d+r+1, which implies that n, the dimension of $\underline{G}$, is greater than or equal to d+r+1.

Theorem 8.8.

Suppose the DLM (4.17) satisfies (8.3). Then the model is observable if and only if n = d+r+1.

Proof

The observability matrix $\left[\underline{F}^T(\underline{FG})^T \ldots (\underline{FG}^{n-1})^T\right]^T$ given in (5.19) has the same rank as $\left[\underline{F}^T \underline{R}_{r,n-r-1}^T\right]^T$, since the rows of one matrix are linear combinations of rows of the other. If the model is observable, then $\left[\underline{F}^T \underline{R}_{r,n-r-1}^T\right]^T$ must have rank n. But the last row of this matrix is $\underline{FG}^{r+1}(\underline{G}-\underline{I})^{n-r-2}$, which from equation (8.3) is zero for $n-r-2 \geq d$, hence if the model is observable, $n<d+r+2$ or $n \leq d+r+1$.

From the corollary to Lemma 8.7, $n \geq d+r+1$, hence $n=d+r+1$.

Conversely, if n=d+r+1, the observability matrix has the same rank as $\left[\underline{F}^T \ \underline{R}_{r,d}^T\right]^T$ which we know to have full rank.

Lemma 8.9

When $n = d+r+1$, $\text{rank}\ (\underline{G}) = d + r$.

Proof

The proof of this result is different from that of Lemma 7.13, only in that $\underline{R}_d$ is replaced by $\underline{R}_{r,d}$. Thus we postmultiply $\left[\underline{F}^T\ \underline{R}_{r,d}^T\right]^T$ by $\underline{G}$, to obtain

$$\begin{bmatrix} \underline{F} \\ \underline{R}_{r,d} \end{bmatrix} \underline{G} = \begin{bmatrix} \underline{0} & \underline{1} & \underline{0} & \cdots & \underline{0} \\ \underline{0} & & & \underline{K} & \end{bmatrix} \begin{bmatrix} \underline{F} \\ \underline{R}_{r,d} \end{bmatrix}$$

where (here) $\underline{K}$ is defined by (8.8). From this,

$$\text{rank}\ (\underline{G}) = \text{rank}\ (\underline{K}) = n-1.$$

It follows that the forward shifted model is not a direct generalisation of the polynomial model. There is only one possible dimension for an observable model, and we have found no general restrictions on the stability region. However, this does not mean that the stability region will necessarily be unrestricted.

Example 8.1

We can verify this point by the following illustrative example with $r = 1$.

Take $\underline{F} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$, $\underline{G} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $\underline{W}$ diagonal

and $v > 0$.

The DLM described by these quantities satisfies $\underline{FG}^2(\underline{G}-\underline{I}) = \underline{0}$, $\underline{FG}^2 \neq \underline{0}$, $\underline{FG}(\underline{G}-\underline{I}) \neq \underline{0}$. In this case, $r = 1$ and $d = 1$. Since $n = 3 = r+d+1$, the model is observable.

Let

$$\underline{C}_{t-1} = \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}$$

Then

$$\underline{P}_t = \underline{G}\underline{C}_{t-1}\underline{G}^T + \underline{W} = \begin{bmatrix} d+w_1 & 0 & e \\ 0 & w_2 & 0 \\ e & 0 & f+w_3 \end{bmatrix}$$

and

$$\underline{A}_t = \underline{P}_t\underline{F}^T/(\underline{F}\underline{P}_t\underline{F}^T + v) = \begin{bmatrix} d + e + w_1 \\ w_2 \\ e + f + w_3 \end{bmatrix} /(d+2e+f+w_1+w_2+w_3+v).$$

The time dependent value of β_2 is given by

$$\beta_2 = \underline{F}\underline{G}(\underline{G}-\underline{I})\underline{A}_t = \begin{bmatrix} 0 & -1 & 0 \end{bmatrix} \underline{A}_t$$

$$= -w_2/(d+2e+f+w_1+w_2+w_3+v).$$

Clearly, $-1 \leq \beta_2 \leq 0$ for all t, thus this model cannot cover the full region.

Example 8.2

As in Example 8.1, $r = 1$, $d = 1$. We take

$$\underline{G} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad \underline{F} = \begin{bmatrix} f_1 & f_2 & f_3 \end{bmatrix}$$

with $\underline{W}$ diagonal, and $\text{var}(\varepsilon_t) = v > 0$.

From Lemma 8.8, the model must be observable, and this implies $f_1 \neq 0$. For simplicity, let

$$\underline{C}_{t-1} = \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}.$$

Then

$$\underline{P}_t = \underline{G}\underline{C}_{t-1}\underline{G}^T + \underline{W} = \begin{bmatrix} d & e & e \\ e & f & f \\ e & f & f \end{bmatrix} + \begin{bmatrix} w_1 & & \\ & w_2 & \\ & & w_3 \end{bmatrix}.$$

Using equations (4.4) to (4.8),

$$\underline{A}_t = \left[A_{1t}\ A_{2t}\ A_{3t}\right]^T = \underline{P}_t\underline{F}^T/(\underline{F}\underline{P}_t\underline{F}^T + \underline{v}) \ .$$

Then

$$A_{1t} = f_1(d+w_1) + f_2e + f_3e/D$$

$$A_{2t} = f_1e + f_2(f+w_2) + f_3f/D$$

$$A_{3t} = f_1e + f_2f + f_3(f+w_3)/D$$

where

$$D = f_1^2d + 2(f_1f_2+f_1f_3)e + (f_2+f_3)^2f + f_1^2w_1+f_2^2w_2+f_3^2w_3 + v \cdot$$

Assuming steady state conditions, so that $\underline{A}_t = \underline{A}$ for all t, and the theory of Section 8.2 is applicable, we have

$$\begin{aligned}\beta_2 &= \underline{F}\underline{G}(\underline{G}-\underline{I})\underline{A} \\ &= \left[0\ -f_1\ f_1\right]\underline{A} \\ &= f_1(A_3 - A_2) \\ &= f_1(f_3w_3 - f_2w_2)/D.\end{aligned}$$

Clearly, if w_2 becomes large compared to the other variances, then

$$\beta_2 \to -f_1/f_2.$$

Similarly, if w_3 becomes large, then

$$\beta_2 \to f_1/f_3.$$

For a flexible DLM, we normally want β_2 to be able to take all values between -1 and 1. Hence either

(a) $\quad f_1 = f_2 = f_3$

or

(b) $\quad f_1 = -f_2 = -f_3.$

First consider(a).

$$\underline{F} = f_1 \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$$

$$\beta_1 = \underline{FGA} - 1$$

$$= f_1 (A_2 + 2A_3) - 1.$$

Hence

$$1 + \beta_1 + \beta_2 = 3f_1A_3$$

$$= 3f_1^2(e + 2f + w_3)/D$$

and

$$1 - \beta_1 + \beta_2 = 2 - 2f_1A_2 - f_1A_3$$

$$= 2 - f_1^2(3e + 6f + 2w_2 + w_3)/D.$$

Now considering(b)

$$F = f_1 \begin{bmatrix} 1 & -1 & -1 \end{bmatrix}$$

so that

$$\beta_2 = f_1^2(w_2 - w_3)/D$$

and

$$\beta_1 = \underline{FGA} - 1 = f_1(A_2 - 2A_3) - 1$$

Hence

$$1 + \beta_1 + \beta_2 = -f_1A_3 = -f_1^2(e-2f-w_3)/D$$

and

$$1 - \beta_1 + \beta_2 = 2 - 2f_1A_2 + 3f_1A_3 = 2 + f_1^2(e-2f+2w_2-3w_3)/D.$$

Unfortunately, in neither case is it possible to draw precise conclusions about the range of β_1 from DLM considerations alone, but either of these cases seems to be a significant improvement on Example 8.1, since β_2 can take the full range of values.

We consider case (a) in more detail by adopting the covariance argument of Harrison (1967) as in Chapter 7. For simplicity, let $f_1 = 1$. We can write the DLM in the form

$$\nabla y_t = \nabla\theta_{1t} + \nabla\theta_{2t} + \nabla\theta_{3t} + \nabla\varepsilon_t$$

$$= \nabla w_{1t} + w_{2t} - w_{2t-2} + w_{3t} + w_{3t-1} + w_{3t-2} + \nabla\varepsilon_t \quad (8.20)$$

which has the same correlation structure as an ARIMA (0,1,2) model

$$\nabla y_t = a_t + \beta_1 a_{t-1} + \beta_2 a_{t-2} \,. \quad (8.21)$$

As in previous examples, we multiple ∇y_t in turn by ∇y_t, ∇y_{t-1}, ∇y_{t-2} and take expectations. Equating the results from equations (8.20) and (8.21), we obtain

$$E\left[(\nabla y_t)^2\right] = 2w_1 + 2w_2 + 3w_3 + 2v = (1 + \beta_1^2 + \beta_2^2)\mathrm{var}(a)$$

$$E\left[(\nabla y_t)(\nabla y_{t-1})\right] = -w_1 + 2w_3 - v = \beta_1(1+\beta_2)\mathrm{var}(a)$$

$$E\left[(\nabla y_t)(\nabla y_{t-2})\right] = -w_2 + w_3 = \beta_2\mathrm{var}(a).$$

Solving these three equation yields

$$w_1 + v = (2(1+\beta_1+\beta_2)^2/9 - \beta_1(1+\beta_2))\mathrm{var}(a)$$

$$w_2 = ((1 + \beta_1 + \beta_2)^2/9 - \beta_2)\mathrm{var}(a)$$

$$w_3 = (1 + \beta_1 + \beta_2)^2\mathrm{var}(a)/9,$$

from which we see that

$$2(1 + \beta_1 + \beta_2)^2 - 9\beta_1(1 + \beta_2) \geq 0$$

and

$$(1 + \beta_1 + \beta_2)^2 - 9\beta_2 \geq 0 \,.$$

$2(1 + \beta_1 + \beta_2)^2 - 9\beta_1(1 + \beta_2) = 0$ can be reduced to the two straight lines

$$2\beta_1 - \beta_2 - 1 = 0 \quad \text{and} \quad \tfrac{1}{2}\beta_1 - \beta_2 - 1 = 0 \,.$$

Thus the stability region covered by this model is given by the shaded area of Figure 8.2.

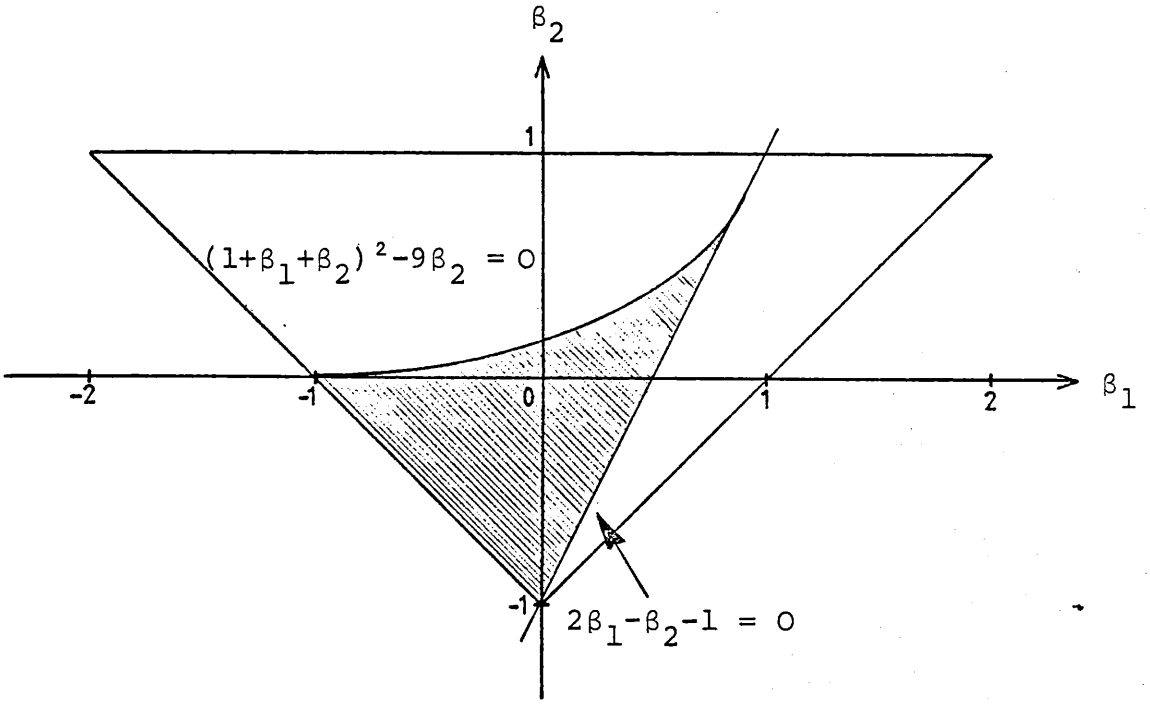


FIGURE 8.2: STABILITY REGION FOR EXAMPLE 8.2.

Example 8.3

Another interesting if inconclusive example is given by

$$y_t = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \underline{\theta}_t + \varepsilon_t$$

$$\underline{\theta}_t = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \underline{\theta}_{t-1} + \underline{w}_t$$

where

$$\underline{W} = E\left[\underline{w}_t \underline{w}_t^T\right] = \begin{bmatrix} w_1 & & \\ & w_2 & \\ & & w_3 \end{bmatrix}$$

Then

$$\underline{FG} = \begin{bmatrix} 3 & 1 & 0 \end{bmatrix}$$

$$\underline{FG}^2 = \begin{bmatrix} 4 & 0 & 0 \end{bmatrix}$$

$$\underline{FG}^2(\underline{G}-\underline{I}) = \underline{0}$$

$$\underline{FG}(\underline{G}-\underline{I}) = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix}.$$

Thus this example represents a forward shifted polynomial model, with $d = 1$, and $r = 1$. $\underline{F}$, $\underline{FG}$, and $\underline{FG}^2$ are linearly independent, hence the model is observable, so that an equilibrium state exists and the theory of Section 8.2 is applicable. Assuming the steady state, with Kalman gain matrix

$$\underline{A} = \begin{bmatrix} A_1 & A_2 & A_3 \end{bmatrix}^T,$$

$$\underline{FGA} = 3A_1 + A_2$$

$$\underline{FG}^2\underline{A} = 4A_1.$$

From equation (8.14),

$$\beta_1 = -1 + \underline{FGA} = 3A_1 + A_2 - 1$$

$$\beta_2 = \underline{FG}^2\underline{A} - \underline{FGA} = A_1 - A_2.$$

Thus

$$1 + \beta_1 + \beta_2 = 4A_1$$

$$1 - \beta_1 + \beta_2 = 2(1 - A_1 - A_2).$$

If the model is to cover the full stability region, then it is required to show

(i) $A_1 > 0$

(ii) $1 - A_1 - A_2 > 0$

(iii) $-1 < A_1 - A_2 < 1$.

Since it is difficult to solve the steady state equations, we shall attempt to find conditions under which (i), (ii) and (iii) are satisfied for all t, and hence also in

the steady state. For simplicity we write

$$\underline{C}_{t-1} = \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}$$

Then using equations (4.4) - (4.7) we find that

$$A_{1t} = (3a + b + w_1)/D_t$$

$$A_{2t} = (3a + b + w_2)/D_t$$

$$A_{3t} = (3a + 4b + d + w_3)/D_t$$

where

$$D_t = 9a + 6b + d + w_1 + w_2 + w_3 + v.$$

Thus the above conditions become

(i) $3a + b + w_1 > 0$

(ii) $3a + 4b + d + w_3 + v > 0$

(iii) $-1 < (w_1 - w_2)/D_t < 1.$

Condition (iii) is clearly satisfied, and all values in the range (-1,1) are possible, since if w_1 is large, then $\beta_2 \to 1$, while if w_2 is large, $\beta_2 \to -1$. Using equation (4.8), it is possible to find expressions for the elements of $\underline{C}_t$ in terms of those of $\underline{C}_{t-1}$.

It turns out that if we set

$$(3a + b + w_1)_o \geq 0$$

$$(3a + b + w_1)_1 \geq 0$$

and

$$(33a + 23b + 4d + 3w_1 + 4(w_2 + w_3 + v))_o \geq 0,$$

then

$$(3a + b + w_1)_t > 0 \quad \text{for all} \quad t.$$

Unfortunately, it appears difficult to draw a positive conclusion about condition (ii). The only covariance involved is b, all other terms are variances and hence positive, so it seems likely that (ii) is satisfied.

Following the same argument as in the previous example, we find that

$$\nabla y_t = w_{1t}+2w_{1t-1}+w_{1t-2}+w_{2t}-w_{2t-2}+\nabla(w_{3t}+\varepsilon_t).$$

This has the same correlation structure as the ARIMA (0,1,2) process

$$\nabla y_t = a_t + \beta_1 a_{t-1}+\beta_2 a_{t-2} \quad .$$

Thus we have

$$E\left[(\nabla y_t)^2\right] = 6w_1+2w_2+2(w_3+v) = (1+\beta_1^2+\beta_2^2)\,var(a)$$

$$E\left[(\nabla y_t)(\nabla y_{t-1})\right] = 4w_1-(w_3+v) = \beta_1(1+\beta_2)\,var(a)$$

$$E\left[(\nabla y_t)(\nabla y_{t-2})\right] = w_1-w_2 = \beta_2\,var(a).$$

Solving these equations

$$16w_1 = (1+\beta_1+\beta_2)^2$$

$$4(w_3+v) = (1-\beta_1+\beta_2)^2$$

$$16w_2 = (1+\beta_1+\beta_2)^2 - 16\beta_2.$$

Thus we see that the stability region is restricted by

$$(1+\beta_1+\beta_2)^2 - 16\beta_2 \geq 0$$

as shown in Figure 8.3.

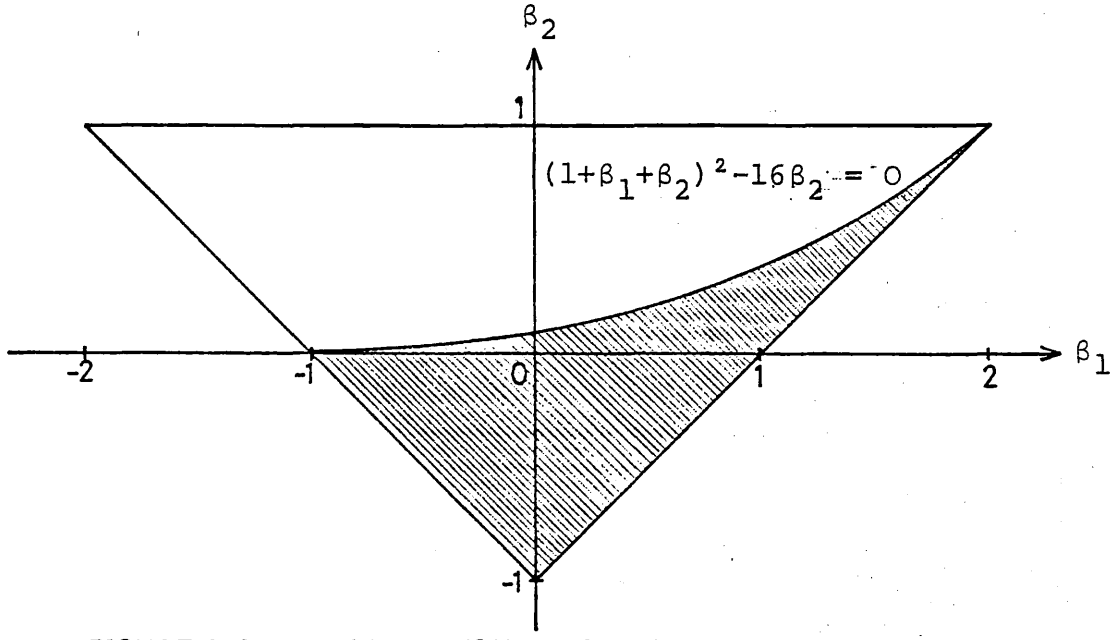


FIGURE 8.3:STABILITY REGION FOR EXAMPLE 8.3.

This is a larger region than that of Example 8.2. Since we have been unable to find conditions for $1-\beta_1+\beta_2 > 0$, it is possible that this model is not stable, i.e. the β's could fall outside the triangular region.

CHAPTER 9

GENERALISATION

9.1 Representation of Model

In this chapter, we move away from systems which yield a polynomial predictor and consider a more general case. The general model is defined by

$$y_t = \underline{F}\underline{\theta}_t + v_t \tag{9.1}$$

$$\underline{\theta}_t = \underline{G}\underline{\theta}_{t-1} + \underline{w}_t . \tag{9.2}$$

As before, y_t is the single observation, $\underline{\theta}_t$ is the $n \times 1$ state vector, v_t, $\underline{w}_t$ are random variables of dimension one and n respectively, with zero means, variance V and covariance matrix $\underline{W}$. $\underline{F}$ is a $1 \times n$ row vector, and $\underline{G}$ an $n \times n$ matrix, both known and independent of t. These matrices will usually be very different from those in the preceeding chapters. In particular, $\underline{G}$ will involve real-valued parameters. In principle, $\underline{G}$ could involve complex-valued parameters, but this situation appears to be unrealistic for practical purposes, and we will always attempt to avoid models of this kind.

Again from Section 4.3 we take the linear predictor of y_{t+k} at time t to be

$$y_t(k) = \underline{F}\underline{G}^k\hat{\underline{\theta}}_t \tag{9.3}$$

where $\hat{\underline{\theta}}_t$ is given by the Kalman updating equation

$$\hat{\underline{\theta}}_t = \underline{G}\hat{\underline{\theta}}_{t-1} + \underline{A}_t(\underline{y}_t - \underline{F}\underline{G}\hat{\underline{\theta}}_{t-1}). \tag{9.4}$$

$\underline{A}_t$ is of the same form as in Chapter 4, namely

$$\underline{A}_t = \underline{P}_t\underline{F}^T/(\underline{F}\underline{P}_t\underline{F}^T + V) \tag{9.5}$$

where

$$\underline{P}_t = \underline{G}\underline{C}_{t-1}\underline{G}^T + \underline{W} \tag{9.6}$$

and

$$\underline{C}_t = E\left[(\underline{\theta}_t - \hat{\underline{\theta}}_t)(\underline{\theta}_t - \hat{\underline{\theta}}_t)^T\right]$$

$$= (\underline{I} - \underline{A}_t\underline{F})\underline{P}_t . \tag{9.7}$$

To generalise the theory of the preceeding chapters in a way which can be useful in practice, it is necessary to make more specific assumptions about some of the parameters of the system. We attempt to achieve this generality by investigating the structure of the matrices $\underline{F}$ and $\underline{G}$, and the following two very general conditions will be assumed.

For some given $r \geq 0$, $s > 0$:-

Assumption 1

The row vectors $\underline{F}\underline{G}, \underline{F}\underline{G}^2, \ldots, \underline{F}\underline{G}^{r+s}$ are linearly independent.

Assumption 2

$\underline{F}\underline{G}^{r+s+1}$ belongs to the vector space spanned by $\underline{F}\underline{G}^{r+1}, \ldots, \underline{F}\underline{G}^{r+s}$.

Assumption 2 implies that there are real constants $\phi_1, \phi_2, \ldots, \phi_s$ such that

$$\underline{F}\underline{G}^{r+s+1} + \phi_1\underline{F}\underline{G}^{r+s} + \ldots + \phi_s\underline{F}\underline{G}^{r+1} = \underline{0} . \tag{9.8}$$

The values assumed by the ϕ_j turn out to be crucial to the basic properties of the model, and it is helpful to introduce the notation $\Phi(z)$ by

$$\Phi(z) = z^s + \phi_1 z^{s-1} + \ldots + \phi_{s-1}z + \phi_s . \tag{9.9}$$

It is suggested here that a useful class of statistical models is derived if $\Phi(z)$ factorises into two polynomials, one with d repeated roots of unity, and the other a polynomial $\alpha(z)$ with all its $p = s-d$ roots strictly inside the unit circle. Accordingly, we make the following assumption about the polynomial (9.9).

Assumption 3

$$\Phi(z) = (z-1)^d \alpha(z) \qquad (9.10)$$

where

$$\alpha(z) = z^p + \alpha_1 z^{p-1} + \ldots + \alpha_p$$

with $p = s-d$, and all roots of $\alpha(z)$ are less than one in modulus.

The motivation for this assumption stems from the comparison of the general DLMs of this chapter with the ARIMA models of Box-Jenkins.

To ensure that steady state theory can be applied, we shall require all the models we use to be observable. This means we make a further assumption.

Assumption 4

The model defined by (9.1) and (9.2) is observable. ie, the n vectors

$$\underline{F}, \underline{F}\underline{G}, \ldots, \underline{F}\underline{G}^{n-1}$$

are linearly independent.

Under assumptions 1,2 and 4, there are constraints on n, the dimension of the system vector $\underline{\theta}_t$, in terms of r and s. We have

Lemma 9.1

Under assumptions 1,2 and 4,

$$r+s \leq n \leq r+s+1$$

Proof

The $r+s$ linearly independent row vectors $\underline{FG}$, $\underline{FG}^2,\ldots,\underline{FG}^{r+s}$ are $1 \times n$, hence the maximum number of these vectors which can be linearly independent is n, thus $s+r \leq n$.

From assumption 4 , $\underline{F}$, $\underline{FG},\ldots\underline{FG}^{n-1}$ are linearly independent. But from assumption 2 , $\underline{FG}^{r+1}$, $\underline{FG}^{r+2},\ldots,\underline{FG}^{r+s+1}$ are linearly dependent. Hence $n-1 < r+s+1$, and the lemma follows.

The three assumptions leading up to Lemma 9.1 have clearly left us in a similar position to that of the previous three chapters, where there is a choice of only two dimensions for the system vector. We now investigate this situation further, with the object of generalising the results obtained for polynomial DLMs. Firstly, we obtain a matrix representation for the relationships between the predictors $y_t(k) = \underline{FG}^k\hat{\underline{\theta}}_t$ of various lead times k, at times t and $t-1$. To this end, we premultiply the Kalman updating equation (9.4) by $\underline{FG}^k (k > 0)$ to obtain

$$\underline{FG}^k\hat{\underline{\theta}}_t = \underline{FG}^{k+1}\hat{\underline{\theta}}_{t-1} + \underline{FG}^k\underline{A}_t(y_t - \underline{FG}\hat{\underline{\theta}}_{t-1}) . \qquad (9.11)$$

If, for simplicity in what follows, we write

$f_{k,t}$ for the k step ahead forecast $\underline{FG}^k\hat{\underline{\theta}}_t$

$\mu_{k,t}$ for the 'smoothing constant' $\underline{FG}^k\underline{A}_t$

and

e_t for the one step ahead prediction error $(y_t - \underline{FG}\hat{\underline{\theta}}_{t-1})$ then equation (9.11) becomes

$$f_{k,t} = f_{k+1,t-1} + \mu_{k,t} e_t . \qquad (9.12)$$

This is valid for all $k > 0$, and in particular for $1 \leq k \leq r+s-1$. When $k = r+s$,

$$\underline{FG}^{r+s} \hat{\underline{\theta}}_t = \underline{FG}^{r+s+1} \hat{\underline{\theta}}_{t-1} + \mu_{r+s,t} e_t .$$

Now using assumption 2 and substituting for $\underline{FG}^{r+s+1}$ from (9.8)

$$\underline{FG}^{r+s} \hat{\underline{\theta}}_t = -(\phi_1 \underline{FG}^{r+s} + \ldots + \phi_s \underline{FG}^{r+1}) \hat{\underline{\theta}}_{t-1} + \mu_{r+s,t} e_t$$

$$f_{r+s,t} = -(\phi_1 f_{r+s,t-1} + \ldots + \phi_s f_{r+1,t-1}) + \mu_{r+s,t} e_t . \qquad (9.13)$$

Combining (9.12) and (9.13), we find that

$$\underline{f}_t = \begin{bmatrix} 0 & 1 & 0 & & & 0 \\ 0 & 0 & 1 & & & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & & & & 1 \\ 0 & \cdot & \cdot & 0 & -\phi_s \cdots & -\phi_1 \end{bmatrix} \underline{f}_{t-1} + \underline{\mu}_t e_t \qquad (9.14)$$

where $\underline{f}_t$, $\underline{\mu}_t$ are the 'prediction' and 'smoothing' vectors respectively, given by

$$\underline{f}_t = \left[f_{1t}, \ldots\ldots\ldots f_{r+s,t} \right]^T$$

$$\underline{\mu}_t = \left[\mu_{1t}, \ldots\ldots\ldots \mu_{r+s,t} \right]^T .$$

Equation (9.14) is compactly represented by

$$\underline{f}_t = \underline{K} \underline{f}_{t-1} + \underline{\mu}_t e_t \qquad (9.15)$$

where $\underline{K} = (k_{ij})$ has elements given by

$$k_{ij} = \begin{cases} 1 & j = i+1 \quad 1 \leq i \leq s+r-1 \\ -\phi_{s+r+1-j} & i = s+r, \quad r+1 \leq j \leq r+s \\ 0 & \text{elsewhere} \end{cases}$$

Equation (9.15) is analagous in many respects to the predictor updating equations of previous chapters, for example, equation (8.8). However, it is worth pointing out that this is not a straightforward generalisation of the polynomial results. In the above, we have effectively defined $\underline{R}_{r+s} = \left[(\underline{FG})^T(\underline{FG}^2)^T \ldots (\underline{FG}^{r+s})^T\right]^T$, but this construction is very different to that of Chapters 7 and 8. For a genuine analogy with the polynomial case, $\underline{R}_{r+s}$ would have involved the roots of $\Phi(z)$, which may be complex. In this case, the form of $\underline{K}$ would have been simpler than that of (9.15), but the matrix would have contained complex elements. Moreover, the matrix $\underline{K}$ will play a role in the construction of a suitable $\underline{G}$ matrix, which would have implied that some elements of $\hat{\underline{\theta}}_t$ are complex linear functions of elements of $\hat{\underline{\theta}}_{t-1}$. This is difficult to interpret realistically when confined to real-valued data. Thus the approach outlined above is considered more appropriate for the general case.

It is interesting to note that the matrix in (9.14) has been considered in a related context by Kalman (1963b); see, for example, the discussion by Priestley (1980) where it is described as one of the canonical forms for a minimal realisation of the linear model

$$X_t + \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} = e_t + \beta_1 e_{t-1} + \ldots + \beta_q e_{t-q}.$$

The characteristic equation of the matrix $\underline{K}$ is

$$\det(\lambda\underline{I} - \underline{K}) = \begin{vmatrix} \lambda & -1 & 0 & & 0 \\ 0 & \lambda & -1 & & 0 \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ \cdot & & & & \cdot \\ 0 & & & \lambda & -1 \\ 0 & 0 & \cdots & \phi_s \cdots \phi_2 & \lambda+\phi_1 \end{vmatrix} = 0 .$$

Expanding by the last row

$$\begin{aligned} \det(\lambda\underline{I}-\underline{K}) &= (\lambda+\phi_1)\lambda^{r+s-1} - \phi_2(-1)\lambda^{r+s-2} + \phi_3(-1)^2\lambda^{r+s-3} \\ &\qquad +\ldots+(-1)^{s+1}\phi_s(-1)^{s-1}\lambda^r \\ &= \lambda^r\left[\lambda^s+\phi_1\lambda^{s-1}+\ldots+\phi_s\right] \\ &= \lambda^r \Phi(\lambda) \end{aligned} \qquad (9.16)$$

where $\Phi(z)$ is defined by equation (9.9). This interesting connection between the predictor updating equation (9.14) and the coefficients of equation (9.8) can be summarised in the following result:

Lemma 9.2

The characteristic equation of the matrix $\underline{K}$ is given by $\lambda^r\Phi(\lambda) = 0$, i.e. $\underline{K}$ has eigenvalues

0 (r times)
1 (d times)

and the p zeros of $\alpha(z)$.

9.2 Steady State Theory

In previous chapters, we found that in the steady state, the DLMs considered were comparable to the ARIMA models described by Box and Jenkins. To find out whether this applies to the general model defined by (9.1) and (9.2), we must assume steady state conditions, so that $\underline{A}_t = \underline{A}$ for all t.

We can now apply z-transforms to equation (9.15), giving

$$\underline{F}(z) = z^{-1}\underline{K}\underline{F}(z) + \underline{\mu}(Y(z) - z^{-1}F_1(z)).$$

In component form, this becomes

$$F_k(z) = z^{-1}F_{k+1}(z) + \mu_k(Y(z) - z^{-1}F_1(z)) \quad 1 \leq k \leq r+s-1$$

$$F_{r+s}(z) = -z^{-1}(\phi_1 F_{r+s}(z)+\ldots+\phi_s F_{r+1}(z)) + \mu_{r+s}(Y(z)-z^{-1}F_1(z)).$$

Rearranging, we have

$$F_{k+1}(z) = zF_k(z)-\mu_k(zY(z)-F_1(z)) \qquad 1 \leq k \leq r+s-1 \quad (9.17)$$

$$zF_{r+s}(z) + \sum_{i=1}^{s} \phi_i F_{r+s+1-i}(z) = \mu_{r+s}(zY(z)-F_1(z)). \quad (9.18)$$

To determine whether or not the model is stable, we need to find the denominator of $F_k(z)/Y(z)$, which, however, is the same for all $k = 1,\ldots,r+s$, since it is given by the determinant of $(z\underline{I}-(\underline{I}-\underline{AF})\underline{G})$, as discussed in Section 5.1. We first prove

Lemma 9.3

$$F_k(z) = F_1(z) \sum_{i=o}^{k-1} \mu_i z^{k-1-i} - Y(z) \sum_{i=1}^{k-1} \mu_i z^{k-i} \quad k = 2,\ldots,r+s \quad (9.19)$$

where $\mu_o = 1$.

Proof (by induction on k)

If we put $k = 1$ in equation (9.17), we find

$$F_2(z) = zF_1(z)-\mu_1(zY(z)-F_1(z))$$

$$= (z+\mu_1)F_1(z)-\mu_1 zY(z).$$

Putting $k = 2$ in equation (9.19) yields

$$F_2(z) = F_1(z)(z+\mu_1)-Y(z)\mu_1 z .$$

This confirms the validity of (9.19) when $k = 2$.

Now suppose that (9.19) is true for some given value of k. Then from equation (9.17),

$$\begin{aligned} F_{k+1}(z) &= zF_k(z)-\mu_k(zY(z)-F_1(z)) \\ &= zF_1(z) \sum_{i=0}^{k-1} \mu_i z^{k-1-i} - zY(z) \sum_{i=1}^{k-1} \mu_i z^{k-i} - \mu_k(zY(z)-F_1(z)) \\ &= F_1(z)\left[\sum_{i=0}^{k-1} \mu_i z^{k-i} + \mu_k\right] - zY(z)\left[\sum_{i=1}^{k-1} \mu_i z^{k-i} + \mu_k\right] \\ &= F_1(z) \sum_{i=0}^{k} \mu_i z^{k-i} - Y(z) \sum_{i=1}^{k} \mu_i z^{k+1-i} \end{aligned}$$

which is (9.19) with k replaced by $k+1$. Since (9.17) is valid for $1 \leq k \leq r+s-1$, equation (9.19) is valid for $2 \leq k \leq r+s$.

Lemma 9.4

The characteristic polynomial of the model is given by

$$z^{r+s} + \sum_{i=1}^{r+s} \mu_i z^{r+s-i} + \sum_{i=1}^{s} \phi_i \sum_{j=0}^{s+r-i} \mu_j z^{s+r-i-j} . \qquad (9.20)$$

Proof

The characteristic polynomial of the model is the denominator of $F_k(z)/Y(z)$, so we need to solve equations (9.17), (9.18) to express $F_k(z)/Y(z)$ in terms of the elements of $\underline{\mu}$.

Substitute for $F_k(z)$ $r+1 \leq k \leq r+s$ from Lemma 9.3 into equation (9.18). This yields

$$zF_{r+s}(z) + \sum_{i=1}^{s} \phi_i F_{r+s+1-i}(z) = \mu_{r+s}(zY(z)-F_1(z))$$

or

$$z\left[F_1(z)\sum_{i=o}^{r+s-1}\mu_i z^{r+s-1-i}-Y(z)\sum_{i=1}^{r+s-1}\mu_i z^{r+s-i}\right]$$

$$+\sum_{i=1}^{s}\phi_i\left[F_1(z)\sum_{j=o}^{r+s-i}\mu_j z^{s+r-i-j}-Y(z)\sum_{j=1}^{s+r-i}\mu_j z^{s+r+1-i-j}\right]$$

$$=\mu_{r+s}(zY(z)-F_1(z)).$$

Collecting terms

$$F_1(z)\left[\sum_{i=o}^{r+s-1}\mu_i z^{r+s-i}+\sum_{i=1}^{s}\phi_i\sum_{j=o}^{r+s-i}\mu_j z^{s+r-i-j}+\mu_{r+s}\right]$$

$$=zY(z)\left[\mu_{r+s}+\sum_{i=o}^{r+s-1}\mu_i z^{r+s-i}+\sum_{i=1}^{s}\phi_i\sum_{j=1}^{s+r-i}\mu_j z^{s+r-i-j}\right].$$

Consequently

$$\frac{F_1(z)}{Y(z)}=\frac{z\left[\sum_{i=o}^{r+s}\mu_i z^{r+s-i}+\sum_{i=1}^{s}\phi_i\sum_{j=1}^{s+r-i}\mu_j z^{s+r-i-j}\right]}{\left[\sum_{i=o}^{r+s}\mu_i z^{r+s-i}+\sum_{i=1}^{s}\phi_i\sum_{j=o}^{r+s-i}\mu_j z^{s+r-i-j}\right]}.$$

The characteristic polynomial of the model is the denominator of this expression, which is the same as (9.20) as required.

The final term of (9.20) can be expressed in a more convenient form as follows:

$$\sum_{i=1}^{s}\phi_i\sum_{j=o}^{r+s-i}\mu_j z^{r+s-i-j}=\sum_{i=1}^{s}\phi_i\sum_{k=i}^{r+s}\mu_{k-i}z^{r+s-k}$$

$$=\sum_{k=1}^{r+s}z^{r+s-k}\sum_{i=1}^{\min(k,s)}\phi_i\mu_{k-i}$$

$$=\sum_{k=1}^{s}z^{r+s-k}\sum_{i=1}^{k}\phi_i\mu_{k-i}$$

$$+\sum_{k=s+1}^{r+s}z^{r+s-k}\sum_{i=1}^{s}\phi_i\mu_{k-i}.$$

Then the characteristic polynomial becomes

$$\beta(z) = z^{r+s} + \sum_{i=1}^{s} z^{r+s-i}\left[\mu_i + \sum_{j=1}^{i} \phi_j \mu_{i-j}\right]$$

$$+ \sum_{i=s+1}^{r+s} z^{r+s-i}\left[\mu_i + \sum_{j=1}^{s} \phi_j \mu_{i-j}\right]$$

or

$$\beta(z) = z^{r+s} + \sum_{i=1}^{s+r} \beta_i z^{r+s-i} \tag{9.21}$$

where

$$\begin{aligned} \beta_i &= \mu_i + \sum_{j=1}^{i} \phi_j \mu_{i-j} \qquad 1 \leq i \leq s \\ \beta_{s+i} &= \mu_{s+i} + \sum_{j=1}^{s} \phi_j \mu_{s+i-j} \qquad 1 \leq i \leq r \,. \end{aligned} \tag{9.22}$$

From Chapter 5, if the model is to be stable, then the zeros of the characteristic polynomial $\beta(z)$ must be less than one in modulus. We are now in a position to prove

Theorem 9.5

Suppose that assumptions 1, 2 and 4 hold, where the zeros of $\Phi(z)$ are less than one in modulus. Then under steady state conditions the k-step ahead predictor $\underline{F}\underline{G}^k\hat{\underline{\theta}}_t$ of the DLM coincides with the k-step ahead predictor of an ARIMA (s,0,r+s) process with moving average parameters $\beta_1, \beta_2, \ldots, \beta_{r+s}$ as defined in (9.22) if and only if the estimation scheme for the DLM is stable in the steady state.

Proof

Equation (9.11) can be rewritten as

$$y_t(k) = y_{t-1}(k+1) + \mu_k e_t \,.$$

So let α_k of Theorem 3.2 be replaced by μ_k. Then the predictor updating equation (9.12) is equivalent to equation (3.20). The theorem then follows from Theorem 3.2.

For if we postmultiply (9.8) by $\underline{G}^{i}\hat{\underline{\theta}}_t$, $i \geq 0$, and then replace $\underline{FG}^{k}\hat{\underline{\theta}}_t$ by $y_t(k)$, we have equation (3.19). The definitions of the β_i given by (3.21), (3.22) and (9.22) are the same, hence Theorem 3.2 can be applied, and the predictors of the DLM are identical to the predictors of the ARIMA (s,0,r+s) model if and only if the roots of $\sum_{i=0}^{r+s} \beta_i z^{r+s-i}$ are less than one in modulus i.e., if and only if the estimation scheme of the DLM is stable.

Corollary

Under assumptions 1, 2, 3 and 4, the k-step ahead predictor $\underline{FG}^{k}\hat{\underline{\theta}}_t$ of the DLM is identical to the k-step ahead predictor of an ARIMA (p,d,p+d+r) process if and only if the estimation scheme is stable in the steady state.

Proof

If $\Phi(z)$ can be factorised into $(z-1)^{d}\alpha(z)$ as in assumption 3, then the proof follows from Theorem 3.3 in the same way as Theorem 9.5 follows from Theorem 3.2.

These two results are the weakest generalisation of the equivalence Theorem 7.6.

9.3 Size and Structure of $\underline{G}$

We already know from Lemma 9.1 that n is either $r+s$ or $r+s+1$. To get a clearer picture of the problem, we first investigate a model with $r=s=1$, which was discussed briefly by Harrison and Stevens (1976) and in the discussion following their paper by Godolphin (1976b).

Example 9.1

$$y_t = u_t + \varepsilon_t$$

$$u_t = g u_{t-1} + \beta_t + \delta\mu_t$$

$$\beta_t = \beta_{t-1} + \delta\beta_t$$

where we assume $|g| < 1$, $g \neq 0$, and $\mathrm{cov}(\delta\mu_t, \delta\beta_t) = 0$. Following the covariance argument as in Section 7.3, we find that

$$\nabla y_t - g\nabla y_{t-1} = \delta\beta_t + \delta\mu_t - \delta\mu_{t-1} + \nabla\varepsilon_t - g\nabla\varepsilon_{t-1}$$

which has the same correlation structure as

$$\nabla y_t - g\nabla y_{t-1} = a_t + \beta_1 a_{t-1} + \beta_2 a_{t-2} .$$

Let $X_t = \nabla y_t - g\nabla y_{t-1}$. Postmultiplying X_t in turn by X_t, X_{t-1} and X_{t-2} and taking expectations, we find that

$$\begin{aligned} E\left[X_t^2\right] &= (1+\beta_1^2+\beta_2^2)\,\mathrm{var}(a) \\ &= \mathrm{var}(\delta\beta) + 2\,\mathrm{var}(\delta\mu) + 2(1+g+g^2)\,\mathrm{var}(\varepsilon) \\ E\left[X_t X_{t-1}\right] &= \beta_1(1+\beta_2)\,\mathrm{var}(a) \\ &= -\mathrm{var}(\delta\mu) - (1+g)^2\,\mathrm{var}(\varepsilon) \\ E\left[X_t X_{t-2}\right] &= \beta_2\,\mathrm{var}(a) \\ &= g\,\mathrm{var}(\varepsilon) . \end{aligned}$$

From these equations, β_2 has the same sign as g, while β_1 is necessarily negative (assuming $\beta_2 > -1$).

$$\mathrm{var}(\delta\mu) = -\left[\beta_1(1+\beta_2) + (1+g)^2\beta_2/g\right]\mathrm{var}(a)$$

which implies that

$$\beta_1(1+\beta_2) + (1+g)^2\beta_2/g \leq 0$$

or

$$\beta_1 \leq -(1+g)^2\beta_2/g(1+\beta_2).$$

The lines

$$\beta_1(1+\beta_2)+(1+g)^2\beta_2/g = 0 \quad \text{and} \quad 1+\beta_1+\beta_2 = 0$$

meet when $\beta_2 = g$, hence the maximum modulus attained by β_2 is $|g|$, thus confirming the conjecture of Godolphin (1976b) that these models are even more restricted than the corresponding polynomial models.

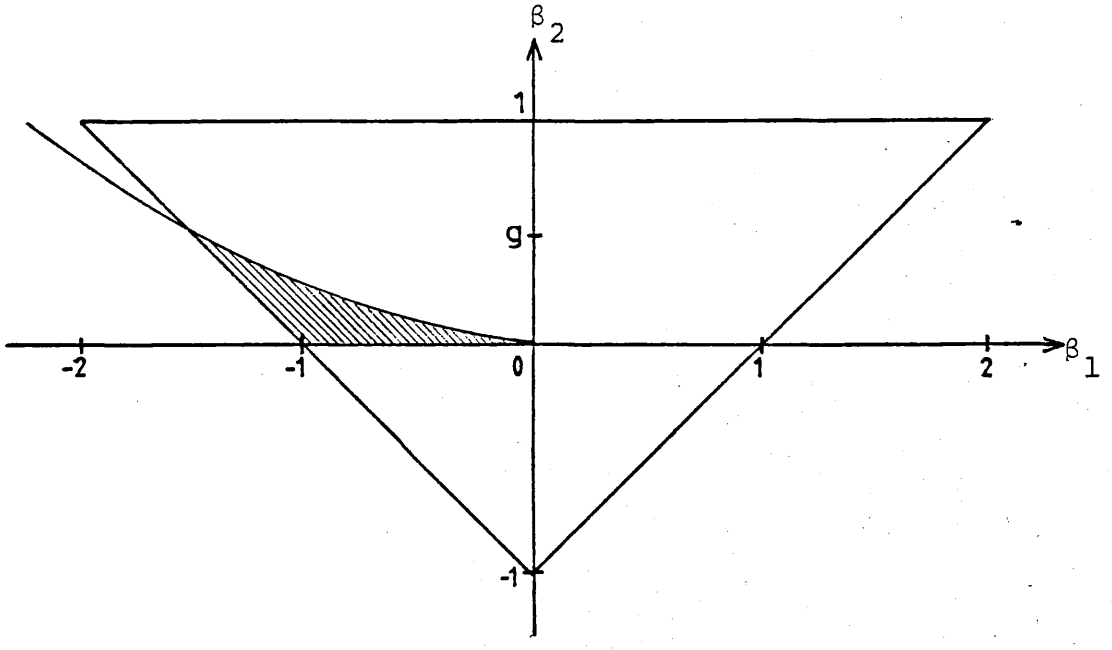


FIGURE 9.1: INVERTIBILITY REGION FOR EXAMPLE 9.1, WITH $0 < g < 1$.

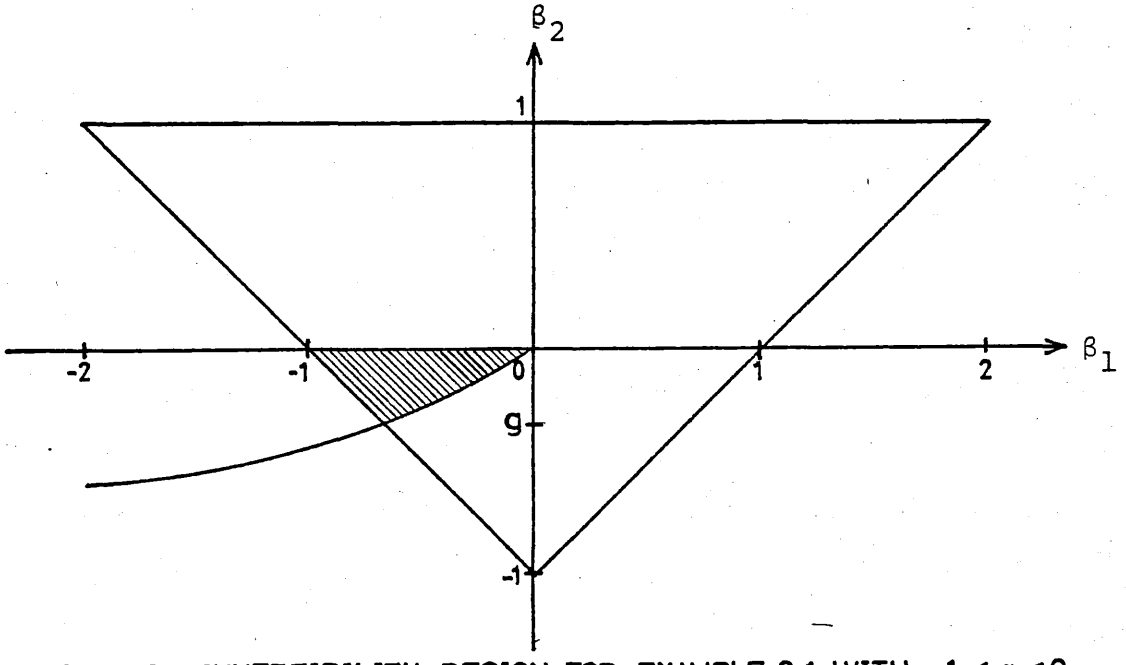


FIGURE 9.2: INVERTIBILITY REGION FOR EXAMPLE 9.1, WITH $-1 < g < 0$.

We seek to enlarge the region in which the parameters can lie. By analogy with previous work, it is reasonable to suppose that the dimension of the model should be enlarged by one. First we prove

Lemma 9.6

Under assumptions 1 , 2 and 4 , the characteristic polynomial of the matrix $\underline{G}$ coincides with the minimal polynomial of $\underline{G}$.

Proof

The minimal polynomial $g(\lambda)$ of any matrix $\underline{A}$ is defined as the monic polynomial of least degree such that $g(\underline{A}) \equiv \underline{0}$. (A monic polynomial is one for which the coefficient of the highest power is one). We first give some results from linear algebra.

1. Any polynomial $h(\lambda)$ for which $h(\underline{A}) \equiv 0$ has a factor $g(\lambda)$.

   For $h(\lambda)$ can be written

   $$h(\lambda) = g(\lambda)a(\lambda) + r(\lambda)$$

   where $r(\lambda)$ has degree less than that of $g(\lambda)$ and $r(\lambda), a(\lambda)$ are unique. Then $h(\underline{A}) \equiv 0$ implies that $r(\underline{A}) \equiv 0$ since $g(\underline{A}) \equiv 0$. But this contradicts the minimality of $g(\lambda)$ unless $r(\lambda) \equiv 0$. Hence $g(\lambda)$ is a factor of $h(\lambda)$.

   We write $g(\lambda) | h(\lambda)$.

2. The minimal polynomial is unique. For if $g_1(\lambda)$, $g_2(\lambda)$ are both minimal polynomials of $\underline{A}$, then

$$g_1(\lambda) | g_2(\lambda) \quad \text{and} \quad g_2(\lambda) | g_1(\lambda),$$

hence $g_1(\lambda)$ is a simple scalar multiple of $g_2(\lambda)$. Since both have leading coefficient unity, we have $g_1(\lambda) = g_2(\lambda)$.

From assumption 4, it is clear that there is no polynomial $g(\lambda)$ of degree less than or equal to $n-1$ such that $g(\underline{G}) = \underline{O}$. Hence the minimal polynomial has degree $\geq n$. But the characteristic polynomial $f(\lambda) = \det(\lambda\underline{I}-\underline{G})$ has degree n and $f(\underline{G}) = \underline{O}$. Hence the minimal polynomial has degree n, and since it is unique, is identical to the characteristic polynomial.

Lemma 9.7

The matrix $\underline{G}$ satisfies

$$s+r \leq \text{rank}(\underline{G}) \leq n .$$

Proof

The linearly independent set $\{\underline{FG}, \underline{FG}^2, \ldots \underline{FG}^{r+s}\}$ can be regarded as a set of mappings on $\underline{G}$, $\{\underline{F}(\underline{G}), \underline{FG}(\underline{G}), \ldots \underline{FG}^{r+s-1}(\underline{G})\}$, hence this is a subset of the image space of $\underline{G}$. The set can be extended to a basis, the dimension of which is defined as the rank of $\underline{G}$. Thus we have $s+r \leq \text{rank}(\underline{G})$. But $\underline{G}$ has dimension n, so that its rank cannot exceed n, and the result follows.

Theorem 9.8

(i) When $n = s+r$, r is necessarily zero, and the characteristic polynomial of $\underline{G}$ is $g(\lambda) = \Phi(\lambda)$.

(ii) when $n = s+r+1$, $g(\lambda) = \lambda^{r+1}\Phi(\lambda)$.

Proof

First we note that if $h(\lambda)$ is any real polynomial such that $\underline{F}h(\underline{G}) = \underline{0}$, then $h(\lambda)$ factorises into $g(\lambda)$ times some other polynomial in λ, i.e. $g(\lambda)|h(\lambda)$. In particular, from equations (9.8) and (9.9)

$$g(\lambda)|\lambda^{r+1}\Phi(\lambda). \tag{9.23}$$

(i) Suppose $n = s+r$, then from Lemma 9.7, rank($\underline{G}$) is n, which implies that zero is not an eigenvalue of $\underline{G}$, and λ is not a factor of the characteristic polynomial of $\underline{G}$. Hence $g(\lambda)|\Phi(\lambda)$, where $g(\lambda)$ has degree n, and $\Phi(\lambda)$ has degree s. It follows that $n = s+r \leq s$, so r must equal zero and the characteristic polynomial is $\Phi(\lambda)$.

(ii) Now suppose $n = s+r+1$. From Lemma 9.7, rank ($\underline{G}$) is n or $n-1$. Using (9.23), the characteristic polynomial $g(\lambda) = \lambda^{r+1}\Phi(\lambda)$ since both are monic of the same degree.

Corollary

(i) When $r > 0$, n must take the value $r+s+1$, the characteristic polynomial is $g(\lambda) = \lambda^{r+1}\Phi(\lambda)$ and rank($\underline{G}$) is $n-1$.

(ii) When $r = 0$, n can be s or $s+1$. When $n = s+1$, $g(\lambda) = \lambda\Phi(\lambda)$, and when $n = s$, $g(\lambda) = \Phi(\lambda)$. In both cases, rank ($\underline{G}$) = s.

For if the characteristic polynomial of $\underline{G}$ is $\lambda^{r+1}\Phi(\lambda)$, then zero is an eigenvalue of $\underline{G}$, so that rank ($\underline{G}$) = $n-1 = s+r$, while if the characteristic polynomial is $\Phi(\lambda)$, then zero is not an eigenvalue, and rank ($\underline{G}$) is n.

By analogy with previous work, it would seem interesting to investigate the range of the β_s when $r = 0$ and $\underline{G}$ is of dimension s. From equations (9.22), we have

$$\beta_s = \mu_s + \sum_{j=1}^{s} \phi_j \mu_{s-j}$$

$$= \underline{F}\underline{G}^s\underline{A} + \sum_{j=1}^{s-1} \phi_j \underline{F}\underline{G}^{s-j}\underline{A} + \phi_s$$

$$= \underline{F}(\underline{G}^s + \sum_{j=1}^{s-1} \phi_j \underline{G}^{s-j})\underline{A} + \phi_s$$

$$= \underline{F}(\underline{G}^s + \sum_{j=1}^{s} \phi_j \underline{G}^{s-j})\underline{A} + \phi_s(1-\underline{F}\underline{A})$$

$$= \underline{F}\Phi(\underline{G})\underline{A} + \phi_s(1-\underline{F}\underline{A}) .$$

When $\underline{G}$ is of dimension s, $\Phi(\lambda)$ is the characteristic polynomial of $\underline{G}$, hence $\Phi(\underline{G}) = \underline{O}$, yielding

$$\beta_s = \phi_s(1-\underline{F}\underline{A}).$$

From equation (9.5), $0 \leq \underline{F}\underline{A} \leq 1$, hence $0 \leq (1-\underline{F}\underline{A}) \leq 1$, thus $|\beta_s|$ lies between 0 and $|\phi_s|$ and β_s takes the sign of ϕ_s. This restriction seems somewhat severe in view of the permissible range for β_s $(-1 < \beta_s < 1)$, and should be avoided if possible. It would therefore appear that an improved dimension for $\underline{G}$ is $n = s+1$. Thus, even when $r = 0$, where there is a choice of dimension for $\underline{G}$, we shall consider only models of dimension $n = r+s+1$. Thus the characteristic polynomial of $\underline{G}$ is $g(\lambda) = \lambda^{r+1}\Phi(\lambda)$, i.e. λ times the characteristic polynomial of $\underline{K}$ (equation (9.16)). This suggests that one choice for a formulation of the matrix $\underline{G}$ is given simply by enlarging the dimension of $\underline{K}$ by one,

remembering that the extra eigenvalue must be zero, and that the rank of $\underline{G}$ must be n-1. Thus

$$\underline{G} = \left[\begin{array}{c|ccc} 0 & 1 & 0 & \cdots & 0 \\ \hline & & & & \\ 0 & & \underline{K} & & \end{array}\right].$$

Example 9.2

Let the characteristic equation of $\underline{G}$ be given by $\lambda(\lambda^2+\phi_1\lambda+\phi_2) = \lambda(\lambda-1)(\lambda+\alpha) = \lambda^3+(\alpha-1)\lambda^2-\alpha\lambda$, where $|\alpha| < 1$. Thus $p=1$, $d=1$ and $r=0$. Then we take

$$\underline{G} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & \alpha & 1-\alpha \end{bmatrix}.$$

We assume $\underline{W}$ is diagonal and positive definite and let $\underline{F} = \begin{bmatrix} f_1 & f_2 & f_3 \end{bmatrix}$. From the observability criterion (5.19), we see that this model is observable if and only if $f_1 \neq 0$. Thus, assuming $f_1 \neq 0$, there is an equilibrium solution, so that the theory of Section 9.2 is applicable. In particular,

$$\beta_2 = -\alpha(1-\underline{FA}) + \underline{F}(\underline{G}+\alpha\underline{I})(\underline{G}-\underline{I})\underline{A}.$$

We define

$$a_t = \text{var}(\theta_{2t}-\hat{\theta}_{2t})$$

$$b_t = \text{cov}((\theta_{2t}-\hat{\theta}_{2t}),(\theta_{3t}-\hat{\theta}_{3t}))$$

$$d_t = \text{var}(\theta_{3t}-\hat{\theta}_{3t}).$$

Then using the time dependent equations (9.5) and (9.6) we find that $\underline{A}_t = \begin{bmatrix} A_{1t} & A_{2t} & A_{3t} \end{bmatrix}^T$ is given by

$$A_{1t} = (a_{t-1}(f_1+\alpha f_3)+b_{t-1}(f_2+(1-\alpha)f_3)+f_1w_1)/D_t$$

$$A_{2t} = (b_{t-1}(f_1+\alpha f_3)+d_{t-1}(f_2+(1-\alpha)f_3)+f_2w_2)/D_t$$

$$A_{3t} = (a_{t-1}\alpha(f_1+\alpha f_3)+b_{t-1}\ (1-\alpha)f_1+\alpha f_2+2\alpha(1-\alpha)f_3$$

$$+d_{t-1}\ (1-\alpha)^2f_3+(1-\alpha)f_2\ +f_3w_3)/D_t$$

where

$$D_t = (f_1+\alpha f_3)^2a_{t-1}+2(f_1+\alpha f_3)(f_2+(1-\alpha)f_3)b_{t-1}$$

$$+(f_2+(1-\alpha)f_3)^2d_{t-1}+f_1^2w_1+f_2^2w_2+f_3^2w_3+v.$$

Therefore, the time dependent value of β_2 is

$$\beta_{2t} = \frac{-\alpha v}{D_t} + f_1(-\alpha A_{1t} + (\alpha-1)\ A_{2t} + A_{3t})$$

$$= (-\alpha v - \alpha f_1^2w_1 + (\alpha-1)f_1f_2w_2 + f_1f_3w_3)/D_t.$$

We would like β_{2t} to be able to take all values in he range $(-1,1)$. We see that

$$\lim_{w_2\to\infty} \beta_{2t} = \frac{(\alpha-1)f_1}{f_2}$$

and

$$\lim_{w_3\to\infty} \beta_{2t} = \frac{f_1}{f_3} \quad .$$

If β_2 is to take all values in the range $(-1,1)$ then either

(a) $\quad f_2 = (\alpha-1)f_1 \quad$ and $\quad f_3 = -f_1$

or

(b) $\quad f_2 = (1-\alpha)f_1 \quad$ and $\quad f_3 = f_1$.

We shall consider the second of these models in more detail.

Example 9.3

Put $f_1 = 1$ in Example 9.2(b). Then the model is given by

$$y_t = \begin{bmatrix} 1 & 1-\alpha & 1 \end{bmatrix} \underline{\theta}_t + \varepsilon_t$$

$$\underline{\theta}_t = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & \alpha & 1-\alpha \end{bmatrix} \underline{\theta}_{t-1} + \underline{w}_t$$

where $|\alpha| < 1$, $\alpha \neq 0$, $\text{var}(\varepsilon_t) = v$ and the covariance matrix of $\underline{w}_t$ is $\begin{bmatrix} w_1 & & \\ & w_2 & \\ & & w_3 \end{bmatrix}$.

Following the same arguments as in Example 9.2, $\underline{A}_t = \begin{bmatrix} A_{1t} & A_{2t} & A_{3t} \end{bmatrix}^T$ is given by

$$A_{1t} = (a_{t-1}(1+\alpha)+2b_{t-1}(1-\alpha)+w_1)/D_t$$

$$A_{2t} = (b_{t-1}(1+\alpha)+2d_{t-1}(1-\alpha)+(1-\alpha)w_2)/D_t$$

$$A_{3t} = (a_{t-1}\alpha(1+\alpha)+b_{t-1}(1-\alpha)(1+3\alpha)+2d_{t-1}(1-\alpha)^2+w_3)/D_t$$

where

$$D_t = (1+\alpha)^2 a_{t-1}+4(1+\alpha)(1-\alpha)b_{t-1}+4(1-\alpha)^2 d_{t-1}+w_1+(1-\alpha)^2 w_2+w_3+v.$$

Thus

$$\beta_{2t} = (-\alpha v-\alpha w_1-(\alpha-1)^2 w_2+w_3)/D_t.$$

Clearly

$$\lim_{w_2 \to \infty} \beta_{2t} = -1$$

and

$$\lim_{w_3 \to \infty} \beta_{2t} = 1 .$$

Thus β_{2t} can take all values in the range $(-1,1)$, as we would expect by the construction of the model from that of Example 9.2. From equation (9.22), the steady state value of β_1 is given by

$$\beta_1 = \underline{FGA}+\phi_1 = \underline{FGA} + \alpha-1$$

so we have

$$1 + \beta_1 + \beta_2 = (3-\alpha)(A_2\alpha+A_3)$$

and

$$1 - \beta_1 + \beta_2 = 2(1-\alpha)-A_2(\alpha^2-\alpha+2)+A_3(3\alpha-1).$$

The time dependent values of these quantities are given by

$$(1+\beta_1+\beta_2)_t = (3-\alpha)\left[a_{t-1}\alpha(1+\alpha)+b_{t-1}(1+3\alpha-2\alpha^2)\right.$$
$$\left.+2d_{t-1}(1-\alpha) + \alpha(1-\alpha)w_2 + w_3\right]/D_t$$

$$(1-\beta_1+\beta_2)_t = \left[a_{t-1}(1+\alpha)(\alpha^2-\alpha+2) + b_{t-1}(5-8\alpha + \alpha^2 - 2\alpha^3)\right.$$
$$+ 2d_{t-1}(1-\alpha)(1-3\alpha)+2(1-\alpha)(w_1+v)$$
$$\left.+ w_2\alpha(1-\alpha)(\alpha-3) + w_3(\alpha+1)\right]/D_t$$

Unfortunately, it is difficult to determine whether these values satisfy $1+\beta_1+\beta_2 > 0$ and $1-\beta_1+\beta_2 > 0$. It is conjectured that these inequalities are satisfied in the steady state.

Assuming the estimation scheme of the model is stable, in the steady state, we follow the covariance argument, as in previous examples.

$$Y_t = \theta_{1t}+(1-\alpha)\theta_{2t}+\theta_{3t}+\varepsilon_t.$$

Let

$$X_t = Y_t+\phi_1 Y_{t-1}+\phi_2 Y_{t-2}$$
$$= Y_t+(\alpha-1)Y_{t-1}-\alpha Y_{t-2}.$$

Writing

$$\psi_{it} = \theta_{it}+(\alpha-1)\theta_{it-1}-\alpha\theta_{it-2} \qquad i = 1,2,3$$

yields

$$X_t = \psi_{1t}+(1-\alpha)\psi_{2t}+\psi_{3t}+\varepsilon_t+(\alpha-1)\varepsilon_{t-1}-\alpha\varepsilon_{t-2}\cdot$$

From the system equations

$$\psi_{1t} = \psi_{2t-1}+w_{1t}+(\alpha-1)w_{1t-1}-\alpha w_{1t-2}$$

$$\psi_{2t} = \psi_{3t-1}+w_{2t}+(\alpha-1)w_{2t-1}-\alpha w_{2t-2}$$

$$\psi_{3t} = \alpha w_{2t-1}+w_{3t}\cdot$$

Writing $u_t = w_{1t} + \varepsilon_t$ some straightforward algebra yields

$$X_t = u_t+(\alpha-1)u_{t-1}-\alpha u_{t-2}+w_{2t}(1-\alpha)+w_{2t-1}(3-\alpha)\alpha$$
$$+(\alpha-1)w_{2t-2}+w_{3t}+(1-\alpha)w_{3t-1}+w_{3t-2}\cdot$$

Thus X_t has the same correlation structure as an MA(2) process

$$X_t = a_t+\beta_1 a_{t-1}+\beta_2 a_{t-2}\cdot$$

Multiplying X_t in turn by X_t, X_{t-1}, X_{t-2} and taking expectations

$$E\left[X_t^2\right] = (1+\beta_1^2+\beta_2^2)\mathrm{var}(a)$$
$$= 2(1-\alpha+\alpha^2)u+(2-4\alpha+11\alpha^2-6\alpha^3+\alpha^4)w_2+(3-2\alpha+\alpha^2)w_3$$

$$E\left[X_tX_{t-1}\right] = \beta_1(1+\beta_2)\mathrm{var}(a)$$
$$= -(1-\alpha)^2u+2(1-\alpha)w_3$$

$$E\left[X_tX_{t-2}\right] = \beta_2\mathrm{var}(a)$$
$$= -\alpha u-(1-\alpha)^2w_2+w_3$$

where $u = \mathrm{var}(w_{1t}+\varepsilon_t)$.

This set of equations implies no obvious restrictions on β_1 and β_2, so it seems possible that this model could cover the complete stability region.

CHAPTER 10

CRAMÉR-RAO BOUNDS

10.1 Discussion

We now turn our attention to a related problem, which is currently of interest to control engineers, namely the application of the Cramér-Rao bound to state-space models. This provides a lower bound on the variance of an estimator, which is clearly useful in determining the quality of the estimators used, and whether a 'better' estimator could be found. This topic seems to have received very little attention in the literature, in particular, the Cramér-Rao bound appears not to have been applied to DLMs. In this chapter, the bound and conditions for it to be attained, will be investigated in the context of the DLMs defined in Chapter 4, where all distributions are assumed to be normal.

However, in contrast to previous work, the system matrix $\underline{G}$ is required to be non-singular. It is usually assumed in the control theory literature that the inverse of $\underline{G}$ exists, although it appears that this assumption is not always necessary. It is possible that the results given below can be extended to a singular $\underline{G}$ by a suitable choice of generalised inverse, but this raises problems of uniqueness. The possibility that $\underline{G}$ may be singular will not be considered here.

We first give a very brief description of Cramér-Rao bound theory.

Suppose $\underline{Y}_1, \underline{Y}_2, \ldots, \underline{Y}_N$ are observed random variables with known joint density function $L(\underline{Y}_1, \underline{Y}_2, \ldots, \underline{Y}_N)$ involving

a vector parameter $\underline{\theta}$. It is required to estimate a function $\underline{a}(\underline{\theta})$ by a function $\underline{t}(\underline{Y}_1,\ldots,\underline{Y}_N)$ of the observations. Then subject to some regularity conditions, the Cramér-Rao bound gives a lower bound on the variance (or covariance matrix) of any unbiased estimator of $\underline{a}(\underline{\theta})$. We shall not consider here the form of the bound for biased estimators. The $\underline{Y}_i$, $\underline{\theta}$, $\underline{a}(\underline{\theta})$, and $\underline{t}(\underline{Y}_1,\ldots,\underline{Y}_N)$ may be vector or scalar valued.

It has been suggested in several texts that the Cramér-Rao bound is attainable if and only if the joint density function of the $\underline{Y}_i$ belongs to the exponential family. A paper of Joshi (1976) gives an example of a non-exponential density for which the Cramér-Rao bound is achieved, but general necessary conditions for attaining the bound seem not to have been established at this time. However, if the density function of the $\underline{Y}_i$ is known to be normal, and the Cramér-Rao lower bound exists, then the bound can be attained under suitable conditions on $\underline{a}(\underline{\theta})$ and its estimator $\underline{t}(\underline{Y}_1,\underline{Y}_2,\ldots,\underline{Y}_N)$.

All the results pertaining to DLMs will assume that the density function of the $\underline{Y}_i$ is multivariate normal. For this reason, the details of Cramér-Rao theory given in the following sections assume that the density function is normal. Thus the likelihood belongs to the exponential family, which enables the order of differentiation and integration to be reversed. This simplifies many calculations, in particular, it means that all but one of the regularity conditions for the Cramér-Rao bound are automatically satisfied.

10.2 Application to the DLM

The DLM has the form

$$\underline{Y}_t = \underline{F}\underline{\theta}_t + \underline{v}_t \tag{10.1}$$

$$\underline{\theta}_t = \underline{G}\underline{\theta}_{t-1} + \underline{w}_t \tag{10.2}$$

where, in general, $\underline{Y}_t$, $\underline{v}_t$ are vectors of dimension m, and $\underline{\theta}_t$, $\underline{w}_t$ are (n x 1) vectors. The $\underline{Y}_t$ are the observations, $\underline{F}$ and $\underline{G}$ are coefficient matrices, assumed known, of dimensions m x n and n x n respectively and $\underline{\theta}_t$ is the unknown 'parameter' to be estimated. $\underline{v}_t$, $\underline{w}_t$ are random variables, each with zero mean and satisfying:

$$E\left[\underline{v}_i\underline{v}_j^T\right] = \underline{V}_i\delta_{ij}, \quad E\left[\underline{w}_i\underline{w}_j^T\right] = \underline{W}_i\delta_{ij}, \quad E\left[\underline{v}_i\underline{w}_j^T\right] = \underline{O} \quad \text{for all} \quad i,j.$$

From equation (10.2), $\underline{\theta}_t$ is a function of a random variable, and hence is a random variable itself, and thus cannot be considered a parameter for the purpose of applying the Cramér-Rao bound to its estimator. However, it is reasonable to consider $\underline{\theta}_1$ a constant (that is, $\underline{\theta}_1 = E(\underline{\theta}_1)$ with probability one), and hence as a parameter. From (10.2)

$$\underline{\theta}_t = \underline{G}^{t-1}\underline{\theta}_1 + \sum_{i=o}^{t-2} \underline{G}^i\underline{w}_{t-i} \tag{10.3}$$

so that $E(\underline{\theta}_t) = \underline{G}^{t-1}\underline{\theta}_1$. For simplicity in what follows the application of the Cramér-Rao bound will be restricted to functions which do not involve random variables, that is, we shall estimate $\underline{\theta}_t$ $(t > 1)$ only when the 'plant noise' is absent. Thus we confine our attention to models of the form

$$\underline{Y}_t = \underline{F}\underline{\theta}_t + \underline{v}_t$$

$$\underline{\theta}_t = \underline{G}\underline{\theta}_{t-1} \cdot \tag{10.4}$$

This enables us to write

$$\underline{y}_t = \underline{F}\underline{G}^{t-1}\underline{\theta}_1 + \underline{v}_t \qquad (10.5)$$

where $\underline{v}_t \sim N(0,\underline{V}_t)$. It is, of course, possible to estimate $\underline{\theta}_1$ whether or not plant noise is present.

It turns out that under these restrictive conditions on the DLM ($\underline{G}$ singular, $\underline{w}_t$ absent, $\underline{v}_t$ normally distributed), the bound can be achieved if and only if the model is observable. This provides us with an additional reason for requiring models to be observable.

10.3 The Scalar Case

Here it is assumed that $m = n = 1$, that is both the observations y_i and the parameter θ are scalar.

The joint probability density function of $y_1,\ldots,y_N$ is denoted by L.

If we wish to estimate $a(\theta)$, then the lower bound on the variance of any unbiased estimator $t(y_1,\ldots,y_N)$ of $a(\theta)$ is given by

$$\text{var}(t) \geq (a'(\theta))^2 / E\left[\left(\frac{\partial \log L}{\partial \theta}\right)^2\right]. \qquad (10.6)$$

The equality in (10.6) is satisfied only when

$$\frac{\partial \log L}{\partial \theta} = f(\theta)t(y_1,\ldots,y_N) + g(\theta) \qquad (10.7)$$

(Rao 1973, p.325), where $f(\theta)$, $g(\theta)$ are functions of θ only.

Since our assumptions imply that the order of differentiation and integration can be reversed,

$$E\left[\left(\frac{\partial \log L}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \log L}{\partial \theta^2}\right] \qquad (10.8)$$

the second expectation being usually much easier to evaluate than the first. For the same reason,

$$E\left[\frac{\partial \log L}{\partial \theta}\right] = 0$$

and by definition, $E(t) = a(\theta)$, hence taking expectations of (10.7) we have

$$g(\theta) = -f(\theta)a'(\theta)$$

so that (10.7) becomes

$$\frac{\partial \log L}{\partial \theta} = f(\theta)(t(y_1,\ldots,y_N) - a(\theta)). \qquad (10.9)$$

Example 10.1

The restricted DLM for the scalar case is given by $y_t = f\theta_t + v_t$ and $\theta_t = g\theta_{t-1}$, where $E[v_t] = 0$, $E[v_t^2] = \sigma_t^2$ and $E[v_t v_j] = 0$, $t \neq j$. Thus $E[y_t] = f\,E[\theta_t] = fg^{t-1}\theta_1$ and

$$E[(y_t - E(y_t))(y_j - E(y_j))] = E[v_t v_j] = \begin{cases} \sigma_t^2 & t=j \\ 0 & t \neq j. \end{cases}$$

Thus the y_i are independent, with mean $fg^{i-1}\theta_1$, variance σ_t^2 and the joint probability density function L of $y_1,\ldots,y_N$ is the product of the density functions of the y_i. If we make the additional assumption that $v_t \sim N(0,\sigma^2)$ for all t, then $y_i \sim N(fg^{i-1}\theta_1,\sigma^2)$ and

$$L = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - fg^{i-1}\theta_1)^2\}.$$

Taking logs, we have

$$\log L = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i^2 - 2fg^{i-1}\theta_1 y_i + (fg^{i-1}\theta_1)^2).$$

In what follows, it is assumed that σ^2 is known. This somewhat unrealistic assumption yields the same results as those obtained when σ^2 is unknown.

We wish to estimate $\theta_N = g^{N-1}\theta_1$, hence

$$a'(\theta_1) = g^{N-1}$$

$$\frac{\partial \log L}{\partial \theta_1} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i - f g^{i-1} \theta_1) f g^{i-1}$$

$$\frac{\partial^2 \log L}{\partial \theta_1^2} = - \frac{1}{\sigma^2} \sum_{i=1}^{N} (f g^{i-1})^2 .$$

Applying equations (10.6) and (10.8), we find that the Cramér-Rao bound for θ_N is given by

$$\text{var}(t) \geq \frac{(g^{N-1})^2 \sigma^2}{\sum_{i=1}^{N} (f g^{i-1})^2}$$

where t is any unbiased estimator of θ_N. Equation (10.9) implies that

$$\sigma^{-2} \sum_{i=1}^{N} (y_i - f g^{i-1} \theta_1) f g^{i-1} = f(\theta_1)(t - g^{N-1} \theta_1).$$

Equating coefficients of θ_1, we find that

$$\sigma^{-2} \sum_{i=1}^{N} (f g^{i-1})^2 = f(\theta_1) g^{N-1}.$$

Thus the estimator of $\theta_N = g^{N-1} \theta_1$ which attains the Cramér-Rao bound is given by

$$t = \frac{1}{f(\theta_1)\sigma^2} \sum_{i=1}^{N} f g^{i-1} y_i = \frac{g^{N-1} \sum_{i=1}^{N} f g^{i-1} y_i}{\sum_{i=1}^{N} f^2 (g^{i-1})^2}$$

$$= \frac{(1-g^2) g^{N-1}}{f(1-g^{2N})} \sum_{i=1}^{N} g^{i-1} y_i \quad \text{for} \quad g \neq 1$$

which is a function of the sufficient statistic $\sum_{i=1}^{N} g^{i-1} y_i$. When $f = g = 1$, this example becomes the familiar steady model. The minimum variance estimator θ_N is then given by

$$t = \sum_{i=1}^{N} y_i / N = \bar{y}$$

with variance σ^2/N.

10.4 Vector parameter θ, scalar observations

In this section, it is assumed that $m = 1$ and n is greater than or equal to one. The generalisation of the Cramér-Rao bound to this vector case, which is described in both the statistical (e.g. Rao, 1973) and engineering (e.g. Eykhoff, 1974) literature, is as follows:

Let $\underline{a}(\underline{\theta}) = \left[a_1(\underline{\theta}),\ldots,a_r(\underline{\theta})\right]^T$ be the vector function of $\underline{\theta}$ to be estimated (in general, $r \leq n$, where n is the dimension of $\underline{\theta}$), and suppose $\frac{\partial \underline{a}}{\partial \underline{\theta}^T}$, which is the $r \times n$ matrix whose (i,j)th element is $\frac{\partial a_i}{\partial \theta_j}$, exists for all $\underline{\theta}$. If $\underline{t}(Y_1,\ldots,Y_N)$ is an unbiased estimator of $\underline{a}(\underline{\theta})$, then subject to some regularity conditions

$$E\left[(\underline{t}-\underline{a})(\underline{t}-\underline{a})^T\right] - \frac{\partial \underline{a}}{\partial \underline{\theta}^T}\, \underline{J}^{-1}\, \frac{\partial \underline{a}^T}{\partial \underline{\theta}} \qquad (10.10)$$

is non-negative definite, where

$$\underline{J} = E\left[\frac{\partial \log L}{\partial \underline{\theta}}\, \frac{\partial \log L}{\partial \underline{\theta}^T}\right] \qquad (10.11)$$

and L is the joint probability density function of $Y_1,\ldots,Y_N$. If as before, it is assumed that L is a member of the exponential family, then although the existence of $\underline{J}^{-1}$ is not guaranteed, all the other regularity conditions are satisfied, and

$$E\left[\frac{\partial \log L}{\partial \underline{\theta}}\, \frac{\partial \log L}{\partial \underline{\theta}^T}\right] = -E\left[\frac{\partial^2 \log L}{\partial \underline{\theta}\, \partial \underline{\theta}^T}\right] . \qquad (10.12)$$

$\frac{\partial \log L}{\partial \underline{\theta}}$ is the $n \times 1$ vector with ith element $\frac{\partial \log L}{\partial \theta_i}$, while the (i,j)th element of the $n \times n$ matrix $\frac{\partial^2 \log L}{\partial \underline{\theta}\, \partial \underline{\theta}^T}$ is given by $\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}$.

Equation (10.10) implies that the generalised variance

$$\det\left(E\left[(\underline{t}-\underline{a})(\underline{t}-\underline{a})^T\right]\right) \geq \det\left(\frac{\partial \underline{a}}{\partial \underline{\theta}^T}\, \underline{J}^{-1}\, \frac{\partial \underline{a}^T}{\partial \underline{\theta}}\right) .$$

Considering the diagonal elements of (10.10) in the case $\underline{a}(\underline{\theta}) = \underline{\theta}$, we have

$$E\left[(t_i - a_i)^2\right] \geq (\underline{J}^{-1})_{ii}.$$

This is greater than or equal to

$$(J_{ii})^{-1} = \left(-E\left[\frac{\partial^2 \log L}{\partial \theta_i^2}\right]\right)^{-1}$$

(Rao, 1973 , p.327) which is the limit in the scalar case.

However, this smaller bound on the variance of an estimator of θ_i was derived for the case θ_j $1 \leq j \leq n$, $i \neq j$ known; if all the θ_j must be estimated, then to achieve this lower bound, the estimator of θ_i must be independent of the other θ_j, otherwise the bound is increased to $(J^{-1})_{ii}$.

We give two conditions for the bound to be attained;

1. If $\frac{\partial \underline{a}}{\partial \underline{\theta}^T}$ is non-singular, then the minimum generalised variance is attained, i.e.

$$\det\left(E\left[(\underline{t}-\underline{a})(\underline{t}-\underline{a})^T\right]\right) = \det\left(\frac{\partial \underline{a}}{\partial \underline{\theta}^T} \underline{J}^{-1} \frac{\partial \underline{a}^T}{\partial \underline{\theta}}\right) \qquad (10.13)$$

if and only if

$$\frac{\partial \log L}{\partial \underline{\theta}} = \underline{C}(\underline{t}-\underline{a}) \qquad (10.14)$$

where $\underline{C}$ is an $n \times n$ matrix independent of the observations (Zacks, 1971, p.195).

2. If $\underline{a}(\underline{\theta}) = \underline{\theta}$, so that $\frac{\partial \underline{a}}{\partial \underline{\theta}^T} = \underline{I}$, then the minimum variance for an estimator of θ_i is attained if and only if

$$t_i - \theta_i = \sum_{j=1}^{n} \lambda_j \frac{\partial \log L}{\partial \theta_j} \qquad (10.15)$$

where the λ_j are independent of the y_i. (Rao, 1945, equation (4.9)).

Clearly, if $E\left[(t_i-\theta_i)^2\right] = (J^{-1})_{ii}$, for some i, then $E\left[(t_i-\theta_i)(t_j-\theta_j)\right] = (J^{-1})_{ij}$ for all j, otherwise the matrix (10.10) is not non-negative definite.

If (10.15) holds for all i, then the matrix (10.10) is zero. That is the complete bound is attained if and only if

$$\underline{t}-\underline{\theta} = \underline{\Lambda} \frac{\partial \log L}{\partial \underline{\theta}} \qquad (10.16)$$

where $\underline{\Lambda}$ is an n x n matrix independent of $y_1,\ldots,y_N$.

It is evidently preferable that the estimator $\underline{t}(y_1,\ldots,y_N)$ has covariance matrix equal to the lower bound, rather than the generalised variance equalling the determinant of the bound. However, when $\underline{a}(\underline{\theta}) \neq \underline{\theta}$, only condition (1) can be applied.

Example 10.2

Using the model given by equations (10.4) with m = 1, let $v_t \sim N(0,\sigma_t^2)$. Then

$$E\left[(y_t-E(y_t))(y_j-E(y_j))\right] = E\left[v_t v_j\right]$$

$$= \begin{cases} \sigma^2_t & t = j \\ 0 & t \neq j \end{cases}$$

so that the y_i are independent and normally distributed. The joint likelihood function of $y_1,\ldots,y_N$ is therefore

$$L = \left(\frac{1}{2\pi}\right)^{N/2} \prod_{i=1}^{N} \sigma_i^{-2} \exp\{-\tfrac{1}{2} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} (y_i - \underline{F}\underline{G}^{i-1}\underline{\theta}_1)^2\}$$

Taking logs, we obtain

$$\log L = \frac{-N}{2} \log 2\pi - \sum_{i=1}^{N} \log \sigma_i^2$$

$$- \frac{1}{2} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} (y_i^2 - 2\underline{\theta}_1^T (\underline{F}\underline{G}^{i-1})^T y_i + \underline{\theta}_1^T (\underline{F}\underline{G}^{i-1})^T \underline{F}\underline{G}^{i-1}\underline{\theta}_1).$$

From the factorisation theorem, $\sum_{i=1}^{N} \frac{1}{\sigma_i^2} (FG^{i-1})^T y_i$ is a sufficient statistic for $\underline{\theta}_1$. Differentiating with respect to $\underline{\theta}_1$,

$$\frac{\partial \log L}{\partial \underline{\theta}_1} = -\tfrac{1}{2} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[-2(\underline{F}\underline{G}^{i-1})^T y_i + 2(\underline{F}\underline{G}^{i-1})^T \underline{F}\underline{G}^{i-1} \underline{\theta}_1 \right]$$

$$\frac{\partial^2 \log L}{\partial \underline{\theta}_1 \partial \underline{\theta}_1^T} = -\sum_{i=1}^{N} \frac{1}{\sigma_i^2} (\underline{F}\underline{G}^{i-1})^T \underline{F}\underline{G}^{i-1}.$$

Hence, using equations (10.11) and (10.12)

$$\underline{J} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} (\underline{F}\underline{G}^{i-1})^T \underline{F}\underline{G}^{i-1}. \qquad (10.17)$$

We wish to estimate $\underline{\theta}_N = \underline{G}^{N-1}\underline{\theta}_1$, hence from (10.10), if $\underline{t}$ is any estimator of $\underline{\theta}_N$,

$$E\left[(\underline{t}-\underline{\theta}_N)(\underline{t}-\underline{\theta}_N)^T \right] - \underline{G}^{N-1}\underline{J}^{-1}(\underline{G}^{N-1})^T \qquad (10.18)$$

is non-negative definite.

Using equation (10.14), the generalised variance is attained if and only if

$$\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[(\underline{F}\underline{G}^{i-1})^T y_i - (\underline{F}\underline{G}^{i-1})^T \underline{F}\underline{G}^{i-1} \underline{\theta}_1 \right] = \underline{C}(\underline{t}-\underline{G}^{N-1}\underline{\theta}_1),$$

where $\underline{C}$ is an $n \times n$ matrix independent of the observations. Equating coefficients of $\underline{\theta}_1$,

$$\underline{J} = \underline{C}\underline{G}^{N-1}$$

hence $\underline{t}$ is given by

$$\underline{t} = \underline{C}^{-1} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} (\underline{F}\underline{G}^{i-1})^T y_i$$

$$= \underline{G}^{N-1}\underline{J}^{-1} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} (\underline{F}\underline{G}^{i-1})^T y_i \qquad (10.19)$$

which is a function of the sufficient statistic. It is easily shown that the variance of $\underline{t}$ given by (10.19) is equal to the Cramér-Rao bound (10.18).

For

$$\underline{t} - E(\underline{t}) = \underline{G}^{N-1}\underline{J}^{-1} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} (\underline{F}\underline{G}^{i-1})^T (y_i - E(y_i)).$$

Thus the covariance matrix of $\underline{t}$

$$E\left[(\underline{t}-E(\underline{t}))(\underline{t}-E(\underline{t}))^T\right] =$$

$$E\left[(\underline{G}^{N-1}\underline{J}^{-1}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}(\underline{FG}^{i-1})^T v_i)((\sum_{i=1}^{N}\frac{1}{\sigma_i^2}(\underline{FG}^{i-1})^T v_i)^T \right.$$

$$\left. \underline{J}^{-1}(\underline{G}^{N-1})^T)\right]$$

$$= \underline{G}^{N-1}\underline{J}^{-1}\, E\left[\sum_{i=1}^{N}\frac{1}{\sigma_i^2}(\underline{FG}^{i-1})^T v_i(\sum_{i=1}^{N}\frac{1}{\sigma_i^2}(\underline{FG}^{i-1})^T v_i)^T\right]\underline{J}^{-1}(\underline{G}^{N-1})^T$$

$$= \underline{G}^{N-1}\,\underline{J}^{-1}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}(\underline{FG}^{i-1})^T E\left[v_i^2\right]\underline{FG}^{i-1}\frac{1}{\sigma_i^2}\,\underline{J}^{-1}(\underline{G}^{N-1})^T$$

$$= \underline{G}^{N-1}\underline{J}^{-1}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}(\underline{FG}^{i-1})^T\sigma_i^2\,\underline{FG}^{i-1}\frac{1}{\sigma_i^2}\,\underline{J}^{-1}(G^{N-1})^T .$$

which from (10.17) reduces to

$$\underline{G}^{N-1}\underline{J}^{-1}(\underline{G}^{N-1})^T$$

which is the Cramér-Rao bound derived in (10.18) as required.

It is easy to see that when $\sigma_i^2 = \sigma^2$ for all i and is unknown, the same results are obtained.

Example 10.3

Consider again the constant velocity model described in Example 4.1, without the plant noise. This is a special case of Example 10.2, with $m = 1$, $n = 2$, $\underline{F} = \begin{bmatrix}1 & 0\end{bmatrix}$, $\underline{G} = \begin{bmatrix}1 & \tau\\ 0 & 1\end{bmatrix}$, and $\sigma_i^2 = \sigma^2$ for all i, where τ is the time between measurements, assumed constant. It can be shown that the sequence of estimators given by the expanding memory α-β filter described in Example 4.1 is identical to that produced by the Kalman filter with zero plant noise, (see Morrison, 1969, Chapter 12). The covariance matrix of the estimator $\underline{\theta}_N$ after N observations is derived in Moon (1977) to be

$$\frac{2\sigma^2}{N(N+1)} \begin{bmatrix} 2N-1 & 3/\tau \\ 3/\tau & 6/(N-1)\tau^2 \end{bmatrix}. \qquad (10.20)$$

Applying equation (10.17), after N measurements

$$\underline{J} = \frac{1}{\sigma^2} \sum_{i=1}^{N} \begin{bmatrix} 1 & (i-1)\tau \end{bmatrix}^T \begin{bmatrix} 1 & (i-1)\tau \end{bmatrix}$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{N} \begin{bmatrix} 1 & (i-1)\tau \\ (i-1)\tau & (i-1)^2\tau^2 \end{bmatrix}$$

$$= \frac{1}{\sigma^2} \begin{bmatrix} N & N(N-1)\tau/2 \\ N(N-1)\tau/2 & N(N-1)(2N-1)\tau^2/6 \end{bmatrix}.$$

Thus

$$\underline{J}^{-1} = \frac{12\sigma^2}{N(N-1)(N+1)\tau^2} \begin{bmatrix} (N-1)(2N-1)\tau^2/6 & -(N-1)\tau/2 \\ -(N-1)\tau/2 & 1 \end{bmatrix}$$

and the lower bound on the covariance matrix of an estimator of $\underline{\theta}_N$ is given by

$$\underline{G}^{N-1}\underline{J}^{-1}(\underline{G}^{N-1})^T = \frac{2\sigma^2}{N(N+1)} \begin{bmatrix} 2N-1 & 3/\tau \\ 3/\tau & 6/(N-1)\tau^2 \end{bmatrix}$$

which is the same as equation (10.20). This indicates that the estimator given by the expanding memory α-β filter for this model attains the Cramér-Rao lower bound.

Using equation (10.19) to find the estimator which achieves the bound, we obtain

$$\underline{t} = \underline{G}^{N-1}\underline{J}^{-1}\sigma^{-2} \sum_{i=1}^{N} (\underline{F}\underline{G}^{i-1})^T y_i$$

$$= \frac{12}{N(N-1)(N+1)\tau^2} \begin{bmatrix} -\frac{(N-1)(N-2)\tau^2}{6} & \frac{(N-1)\tau}{2} \\ -\frac{(N-1)\tau}{2} & 1 \end{bmatrix} \sum_{i=1}^{N} \begin{bmatrix} y_i \\ (i-1)\tau y_i \end{bmatrix}$$

$$= \frac{2}{N(N+1)} \begin{bmatrix} -(N-2) & 3/\tau \\ -3/\tau & 6/(N-1)\tau^2 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \tau \sum_{i=1}^{N} (i-1)y_i \end{bmatrix}$$

$$= \frac{2}{N(N+1)} \begin{bmatrix} -(N+1)\sum_{i=1}^{N} y_i + 3\sum_{i=1}^{N} iy_i \\ \frac{3}{\tau(N-1)}\left[-(N+1)\sum_{i=1}^{N} y_i + 2\sum_{i=1}^{N} iy_i\right] \end{bmatrix}.$$

We shall return to this example, to verify that this value of t does indeed coincide with the estimator of the Kalman filter.

10.5 Vector θ, vector observations

Up to this point, attention has been confined to scalar observations. However, there are many occasions when more than one measurement is taken at one time. One frequently used model is the coupled Kalman filter, where position and velocity in both the x and y directions are estimated from measurements of position in both directions. Often these 'measurements' are calculated from actual measurements of range and bearing, hence the covariance matrix for the observation noise is not diagonal.

For air traffic control, measurements of range, bearing and elevation may be used, from which measurements in the x, y and z directions can be calculated. Again, the covariance matrix for the observation error will not usually be diagonal.

Thus in this section, it is assumed that m and n are both greater than one. The model is defined by (10.4), and the Cramér-Rao bound is given by (10.10) and (10.11) exactly as for scalar observations, but the joint likelihood of the y_i is more difficult to find for vector observations. We shall assume immediately that

$$\underline{v}_t \sim N(\underline{0}, \underline{V}_t), \quad E\left[\underline{v}_t \underline{v}_j^T\right] = \underline{0} \quad t \neq j.$$

Then the $\underline{Y}_t$ are independently distributed as multivariate normal

$$\underline{Y}_t \sim N(\underline{FG}^{t-1}\underline{\theta}_1, \underline{V}_t)$$

and the likelihood function of $\underline{Y}_1, \underline{Y}_2, \ldots, \underline{Y}_N$ is given by

$$L = \prod_{i=1}^{N} \left[(\tfrac{1}{2\pi})^{m/2} (\det \underline{V}_i)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(\underline{Y}_i - \underline{FG}^{i-1}\underline{\theta}_1)^T \underline{V}_i^{-1} (\underline{Y}_i - \underline{FG}^{i-1}\underline{\theta}_1)\} \right].$$

In this case, it is necessary to assume that the covariance matrices $\underline{V}_t = E\left[\underline{v}_t \underline{v}_t^T\right]$ are known for all t.

Taking logs we obtain

$$\log L = \sum_{i=1}^{N} \left[-\frac{m}{2} \log 2\pi - \tfrac{1}{2}\log(\det \underline{V}_i) \right.$$

$$\left. - \frac{1}{2} (\underline{Y}_i^T \underline{V}_i^{-1} \underline{Y}_i - 2\underline{\theta}_1^T (\underline{FG}^{i-1})^T \underline{V}_i^{-1} \underline{Y}_i + \underline{\theta}_1^T (\underline{FG}^{i-1})^T \underline{V}_i^{-1} \underline{FG}^{i-1} \underline{\theta}_1) \right]$$

from which $\sum_{i=1}^{N} (\underline{FG}^{i-1})^T \underline{V}_i^{-1} \underline{Y}_i$ is a sufficient statistic for $\underline{\theta}_1$, by the factorisation theorem. Differentiating with respect to $\underline{\theta}_1$,

$$\frac{\partial \log L}{\partial \underline{\theta}_1} = -\frac{1}{2} \sum_{i=1}^{N} \left[-2(\underline{FG}^{i-1})^T \underline{V}_i^{-1} \underline{Y}_i + 2(\underline{FG}^{i-1})^T \underline{V}_i^{-1} \underline{FG}^{i-1} \underline{\theta}_1 \right]$$

$$\frac{\partial^2 \log L}{\partial \underline{\theta}_1 \partial \underline{\theta}_1^T} = -\sum_{i=1}^{N} (\underline{FG}^{i-1})^T \underline{V}_i^{-1} \underline{FG}^{i-1} .$$

Hence from (10.11) and (10.12), $\underline{J}$ is given by

$$\underline{J} = \sum_{i=1}^{N} (\underline{FG}^{i-1})^T \underline{V}_i^{-1} \underline{FG}^{i-1} . \qquad (10.21)$$

If the function of $\underline{\theta}_1$ to be estimated is $\underline{\theta}_N = \underline{G}^{N-1}\underline{\theta}_1$, then $\frac{\partial \underline{a}}{\partial \underline{\theta}_1^T} = \underline{G}^{N-1}$, and from (10.10), if $\underline{t}$ is any estimator of $\underline{\theta}_N$, then

$$E\left[(\underline{t} - \underline{\theta}_N)(\underline{t} - \underline{\theta}_N)^T\right] - \underline{G}^{N-1}\underline{J}^{-1}(\underline{G}^{N-1})^T \qquad (10.22)$$

is non-negative definite.

From equation (10.14), the minimum generalised variance is attained if $\underline{t}$ satisfies

$$\frac{\partial \log L}{\partial \underline{\theta}_1} = \underline{C}(\underline{t} - \underline{G}^{N-1}\underline{\theta}_1)$$

where $\underline{C}$ is an n x n matrix independent of the observations. Equating coefficients of $\underline{\theta}_1$, we have

$$\sum_{i=1}^{N} (\underline{FG}^{i-1})^T \underline{V}_i^{-1} \underline{FG}^{i-1} = \underline{J} = \underline{CG}^{N-1} .$$

Thus $\underline{C}^{-1} = \underline{G}^{N-1}\underline{J}^{-1}$ and

$$\underline{t} = \underline{G}^{N-1}\underline{J}^{-1} \sum_{i=1}^{N} (\underline{FG}^{i-1})^T \underline{V}_i^{-1} \underline{Y}_i \qquad (10.23)$$

which is a function of the sufficient statistic. Again, it is easily verified that the estimator given by (10.23) achieves the Cramér-Rao bound, that is (10.22) is the zero matrix.

Example 10.4

Consider the very simple example given if the observations Y_{1t}, Y_{2t} at time t are measurements of the distance of some stationary object in the x_1 and x_2 directions subject to random measurement errors v_{1t}, v_{2t}. Mathematically, this is

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} + \begin{bmatrix} v_{1t} \\ v_{2t} \end{bmatrix} .$$

It is assumed that the measurement error $\begin{bmatrix} v_{1t} & v_{2t} \end{bmatrix}^T$ is distributed as multivariate normal, with mean zero and covariance matrix $\underline{V}$ for all t. The state equation is represented by

$$\begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} = \begin{bmatrix} x_{1t-1} \\ x_{2t-1} \end{bmatrix} ,$$

that is, $\underline{F} = \underline{G} = \underline{I}_2$. Hence from (10.21)

$$\underline{J} = \sum_{i=1}^{N} \underline{V}^{-1} = N\underline{V}^{-1}$$

and from (10.22)

$$E\left[(\underline{t}-\underline{\theta}_N)(\underline{t}-\underline{\theta}_N)^T\right] - N^{-1}\underline{V}$$

is non-negative definite, where $\underline{t}$ is any unbiased estimator of $\underline{\theta}_N$. Assuming $\underline{v}_t$ is normally distributed, $\underline{t}$ is defined by equation (10.23) as the estimator

$$\underline{t} = N^{-1}\underline{V} \sum_{i=1}^{N} \underline{V}^{-1} \begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix} = \frac{1}{N} \begin{bmatrix} \sum_{i=1}^{N} y_{1i} \\ \sum_{i=1}^{N} y_{2i} \end{bmatrix}.$$

Comparing the results of this section with those of Section 10.4, and in particular with Example 10.2 (which merely assumes normality), it is clear that equations (10.17), (10.18) and (10.19) are special cases of (10.21), (10.22) and (10.23) respectively. Example 10.1 is a further special case.

All further results will be derived in time of the general vector model of this section.

10.6 Properties of models for which the Cramér-Rao bound is achieved

Assuming $\underline{v}_t$ is normally distributed, $\underline{t}$ is defined by equation (10.23) as the estimator of $\underline{\theta}_N$ for the model (10.4) after N observations, and has covariance matrix $\underline{J}$ defined by (10.21). For clarity these quantities will be denoted by $\underline{t}_N$ and $\underline{J}_N$ respectively.

Theorem 10.1

There is a unique unbiased estimator of $\underline{\theta}_N$ based on N observations whose variance achieves the Cramér-Rao

bound if and only if the model (10.4) is observable, where all the distributions are assumed to be normal.

Proof

The inverse of $\underline{J}_N$ is required both to find the Cramér-Rao bound, and for the estimator $\underline{t}_N$ which achieves the bound. Thus there is a unique estimator which achieves the bound if and only if $\underline{J}_N^{-1}$ exists.

It is clear from the definition (5.24) that the matrix $\underline{J}_N$ defined by (10.21) is the observability matrix for the model (10.4). The model is observable if and only if the observability matrix (5.24) is non-singular, hence result.

Theorem 10.1 shows in a separate context from previous results that observability is a useful property of DLMs.

It will now be shown that $\underline{t}_N = \hat{\underline{\theta}}_N$ and its covariance matrix $\underline{G}^{N-1}\underline{J}_N^{-1}(\underline{G}^{N-1})^T$ defined by equations (10.23) and (10.22) respectively satisfy the Kalman updating equations (4.4) - (4.8). Thus if the estimator of $\underline{\theta}_k$ after k measurements is given by (10.23), and hence has a covariance matrix equal to the Cramér-Rao lower bound, then for all $i \geq k$, the estimator $\underline{t}_i$ of $\underline{\theta}_i$ after i measurements produced by the Kalman filter is an unbiased estimator of $\underline{\theta}_i$ with covariance matrix equal to the Cramér-Rao lower bound.

Theorem 10.2

Given the model (10.4) where $\underline{v}_k \sim N(\underline{0},\underline{V}_k)$ and $\underline{V}_k$ is positive definite, let $\underline{t}_k$ given by (10.23) be an unbiased estimator of $\underline{\theta}_k$ for some $k \geq 1$, and let $\underline{t}_k$ have covariance matrix

$\underline{Q}_k = \underline{G}^{k-1}\underline{J}_k^{-1}(\underline{G}^{k-1})^T$ where $\underline{J}_k$ is defined by (10.21). Then $\underline{t}_j$, $\underline{Q}_j$ satisfy the Kalman updating equations (4.4)-(4.8) for all $j \geq k$.

Proof

The Kalman updating procedure is given by

$$\hat{\underline{\theta}}_k = \underline{G}\hat{\underline{\theta}}_{k-1} + \underline{A}_k(\underline{y}_k - \underline{F}\underline{G}\hat{\underline{\theta}}_{k-1})$$

where

$$\underline{A}_k = \underline{C}_k\underline{F}^T\underline{V}_k^{-1}$$

$$\underline{C}_k = (\underline{I}-\underline{A}_k\underline{F})\underline{P}_k$$

and

$$\underline{P}_k = \underline{G}\underline{C}_{k-1}\underline{G}^T .$$

For some $k \geq 1$, let $\underline{t}_k = \hat{\underline{\theta}}_k$ and $\underline{Q}_k = \underline{C}_k$. From the definition of $\underline{Q}_k$, $\underline{Q}_{k+1}$ is given by

$$\begin{aligned}\underline{Q}_{k+1} &= \underline{G}^k\underline{J}_{k+1}^{-1}(\underline{G}^k)^T \\ &= \underline{G}^k\left[\underline{J}_k + (\underline{G}^k)^T\underline{F}^T\underline{V}_{k+1}^{-1}\underline{F}\underline{G}^k\right]^{-1}(\underline{G}^k)^T \\ &= \left[(\underline{G}^{-k})^T\underline{J}_k\underline{G}^{-k} + \underline{F}^T\underline{V}_{k+1}^{-1}\underline{F}\right]^{-1}.\end{aligned}$$

From the Kalman updating equations

$$\underline{C}_{k+1} = (\underline{I} - \underline{C}_{k+1}\underline{F}^T\underline{V}_{k+1}^{-1}\underline{F})\underline{P}_{k+1} \qquad (10.24)$$

so that

$$\underline{C}_{k+1}(\underline{I} + \underline{F}^T\underline{V}_{k+1}^{-1}\underline{F}\underline{P}_{k+1}) = \underline{P}_{k+1}$$

and hence

$$\underline{C}_{k+1} = \underline{P}_{k+1}(\underline{I} + \underline{F}^T\underline{V}_{k+1}^{-1}\underline{F}\underline{P}_{k+1})^{-1}.$$

This can be written

$$\begin{aligned}\underline{C}_{k+1} &= ((\underline{I} + \underline{F}^T\underline{V}_{k+1}^{-1}\underline{F}\underline{P}_{k+1})\underline{P}_{k+1}^{-1})^{-1} \\ &= (\underline{P}_{k+1}^{-1} + \underline{F}^T\underline{V}_{k+1}^{-1}\underline{F})^{-1}\end{aligned}$$

and since $\underline{Q}_k = \underline{C}_k$, so that $\underline{P}_{k+1} = \underline{G}^k \underline{J}_k^{-1} (\underline{G}^k)^T$, this expression for $\underline{C}_{k+1}$ is identical to that for $\underline{Q}_{k+1}$.

From the Kalman updating equations,

$$\hat{\underline{\theta}}_{k+1} = (\underline{I} - \underline{A}_{k+1}\underline{F})\underline{G}\hat{\underline{\theta}}_k + \underline{A}_{k+1}\underline{Y}_{k+1}$$

$$= (\underline{I} - \underline{C}_{k+1}\underline{F}^T\underline{V}_{k+1}^{-1}\underline{F})\underline{G}\hat{\underline{\theta}}_k + \underline{C}_{k+1}\underline{F}^T\underline{V}_{k+1}^{-1}\underline{Y}_{k+1}.$$

Since $\underline{t}_k = \hat{\underline{\theta}}_k$, substitute for $\hat{\underline{\theta}}_k$ from (10.23) to obtain

$$\hat{\underline{\theta}}_{k+1} = (\underline{I} - \underline{C}_{k+1}\underline{F}^T\underline{V}_{k+1}^{-1}\underline{F})\underline{G}^k\underline{J}_k^{-1} \sum_{i=1}^{k} (\underline{F}\underline{G}^{i-1})^T\underline{V}_i^{-1}\underline{Y}_i + \underline{G}^k\underline{J}_{k+1}^{-1}(\underline{G}^k)^T\underline{F}^T\underline{V}_{k+1}^{-1}\underline{Y}_{k+1}$$

$$= (\underline{I} - \underline{C}_{k+1}\underline{F}^T\underline{V}_{k+1}^{-1}\underline{F})\underline{P}_{k+1}(\underline{G}^{-k})^T \sum_{i=1}^{k} (\underline{F}\underline{G}^{i-1})^T\underline{V}_i^{-1}\underline{Y}_i + \underline{G}^k\underline{J}_{k+1}^{-1}(\underline{G}^k)^T\underline{F}^T\underline{V}_{k+1}^{-1}\underline{Y}_{k+1}.$$

Using equation (10.24), this becomes

$$\hat{\underline{\theta}}_{k+1} = \underline{C}_{k+1}(\underline{G}^{-k})^T \sum_{i=1}^{k} (\underline{F}\underline{G}^{i-1})^T\underline{V}_i^{-1}\underline{Y}_i + \underline{G}^k\underline{J}_{k+1}^{-1}(\underline{G}^k)^T\underline{F}^T\underline{V}_{k+1}^{-1}\underline{Y}_{k+1}$$

and since $\underline{C}_{k+1} = \underline{Q}_{k+1} = \underline{G}^k\underline{J}_{k+1}^{-1}(\underline{G}^k)^T$, we have

$$\hat{\underline{\theta}}_{k+1} = \underline{G}^k\underline{J}_{k+1}^{-1} \sum_{i=1}^{k} (\underline{F}\underline{G}^{i-1})^T\underline{V}_i^{-1}\underline{Y}_i + \underline{G}^k\underline{J}_{k+1}^{-1}(\underline{G}^k)^T\underline{F}^T\underline{V}_{k+1}^{-1}\underline{Y}_{k+1}$$

$$= \underline{G}^k\underline{J}_{k+1}^{-1} \sum_{i=1}^{k+1} (\underline{F}\underline{G}^{i-1})^T\underline{V}_i^{-1}\underline{Y}_i,$$

which is the expression for $\underline{t}_{k+1}$ given by (10.23). Thus by induction, $\underline{t}_i$, $\underline{Q}_i$ satisfy the Kalman updating equations for all $i \geq k$.

Example 10.5

Reconsider the model of Example 10.3

$$Y_t = x_t + v_t$$

$$x_t = x_{t-1} + \tau\dot{x}_{t-1}$$

$$\dot{x}_t = \dot{x}_{t-1} .$$

Let the estimator $\underline{t}_2$ of $\begin{bmatrix} x_2 \\ \dot{x}_2 \end{bmatrix}$ after two measurements be given by

$$\underline{t}_2 = \begin{bmatrix} y_2 \\ (y_2-y_1)/\tau \end{bmatrix} .$$

This estimator has covariance matrix

$$\sigma^2 \begin{bmatrix} 1 & 1/\tau \\ 1/\tau & 2/\tau^2 \end{bmatrix} .$$

Comparison with Example 10.3 shows that this estimator must achieve the Cramér-Rao bound.

Applying the Kalman updating equations,

$$\underline{A}_3 = \begin{bmatrix} 5/6 \\ 1/2\tau \end{bmatrix} , \quad \underline{C}_3 = \sigma^2 \begin{bmatrix} 5/6 & 1/2\tau \\ 1/2\tau & 1/2\tau^2 \end{bmatrix}$$

and

$$\underline{t}_3 = \begin{bmatrix} 1 & \tau \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_2 \\ \frac{y_2-y_1}{\tau} \end{bmatrix} + \begin{bmatrix} 5/6 \\ 1/2\tau \end{bmatrix} (y_3-2y_2+y_1)$$

$$= \begin{bmatrix} (5y_3+2y_2-y_1)/6 \\ (y_3-y_1)/2\tau \end{bmatrix} .$$

As expected from Theorem 10.2, these are exactly the values obtained by setting $k=3$ in the expressions for $\underline{t}_k$ and $G^{k-1}J_k^{-1}(G^{k-1})^T$ in Example 10.3.

Thus the estimators produced by the Kalman filter have been shown to achieve the Cramér-Rao bound when there is no plant noise, provided the information matrix $\underline{J}$ is non-singular, and this is coincident with the observability of the model.

CHAPTER 11

SUMMARY

The purpose of this work is to examine structural properties of the dynamic linear models (DLMs) proposed by Harrison and Stevens (1976). In these models, the observations $\{y_t\}$ are described by the observation equation given by

$$\underline{Y}_t = \underline{F}\underline{\theta}_t + \underline{v}_t,$$

where $\underline{\theta}_t$ has the Markovian representation given by the system equation

$$\underline{\theta}_t = \underline{G}\underline{\theta}_{t-1} + \underline{w}_t.$$

Properties of the DLMs are investigated chiefly in relation to the predictors for univariate time series, and also with regard to minimum-variance estimators. Techniques employed for this purpose are those of statistical time series analysis and modern control theory, which are described in Chapters 2-5. In particular, considerable use is made of the unifying concept of observability and its implications.

In Chapter 6, the constant forecast model is considered. It is pointed out that the intuitive steady DLM in its equilibrium state is equivalent to a proper subset of the class of ARIMA (0,1,1) models, in the sense that their predictors are equivalent. The same steady model is generalised by increasing the dimension of the system vector $\underline{\theta}_t$ to two, and then the model may be equivalent to any ARIMA (0,1,1) model, depending on the

values of the system error covariance matrix $\underline{W} = E\left[\underline{w}_t \underline{w}_t^T\right]$. Furthermore, it is shown that whatever the dimension of $\underline{\theta}_t$, the constant forecast DLM is equivalent in the steady state to any ARIMA (0,1,1) model, provided $\underline{FG} \neq \underline{F}$. However, this result may require the use of covariances in the system error covariance matrix. It is often required that this matrix is diagonal. Several examples of 2 x 2 DLMs with diagonal covariance matrices are investigated, and one of these is found to be equivalent to the complete class of ARIMA (0,1,1) models. Observability is invoked to show that the ideal dimension of $\underline{\theta}_t$ for the constant forecast model is two. It follows that the constant forecast model is a significant generalisation of the ARIMA (0,1,1) models. Advantages of the DLM over the classical model include the ability to predict from little data, to make good use of prior intelligence and subjective information, to take into account changes in the variances of the observations, and also to use unequally spaced observations.

Chapter 7 generalises these ideas to the polynomial model of degree greater than zero, say d-1. Here it is shown that the invertibility of an ARIMA (0,d,q) model for $q \leq d$ is equivalent to the stability of the estimation scheme for the DLM in the equilibrium state. Hence if the DLM is stable in the steady state, then its predictors are identical to those of the ARIMA (0,d,q) model. However, it turns out that if $\underline{G}$ is non-singular, then only a subset of the ARIMA (0,d,q) models are admissible. This conclusion, together with the observability requirement yields a necessary

condition for the DLM to be equivalent to all such ARIMA models viz. that n, the dimension of $\underline{G}$, should be equal to $d+1$. Again, the examples show that the practical requirement that $\underline{W}$ be diagonal may restrict the equivalence to a subset of these ARIMA models.

Chapter 8 considers a further generalisation of the DLMs to forward shifted polynomial models. It turns out that there is no DLM which models this situation for which $\underline{G}$ is non-singular. It is shown that the minimum dimension of $\underline{G}$ for these models is $d+r+1$, where d is the degree of the polynomial and r is the shift, and the observability requirement implies that this value is the maximum dimension. Again, the stability of the estimation scheme for the DLM in the equilibrium state is shown to be equivalent to the invertibility of an ARIMA $(0,d,d+r)$ model.

Chapter 9 generalises these models to derive DLMs equivalent to the ARIMA (p,d,q) models. Thus under the conditions described, the predictors of the DLMs in the equilibrium state are shown to be equivalent to the general ARIMA (p,d,q) models. It is shown that some of these DLMs are restricted to a subset of ARIMA models, but many of these restrictions are avoided if $\underline{G}$ is singular. With observability, this result implies that $n = p+1+\max(d,q)$. In addition, the eigenvalues of $\underline{G}$ are related directly to the parameters on the left-hand side of the corresponding ARIMA model equation. Hence the 'autoregressive' model parameters $\phi_1,\ldots,\phi_p$ are an implicit part of the equivalent DLM. This is in contrast to all preceeding results which specify the parameters

on the right-hand side of the ARIMA model equation purely in terms of the elements of $\underline{F}$, $\underline{G}$, $\underline{W}$ and $\underline{V} = E\left[\underline{v}_t \underline{v}_t^T\right]$. Unfortunately, it appears to be difficult to show for these more general models that the predictor of the DLM is identical in the equilibrium state to the predictor of any ARIMA (p,d,q) model, provided $\underline{W}$ is not constrained to be diagonal, although the conjecture seems reasonable. Indeed, the fact that the stability conditions for the estimation scheme of the DLM are equal to the invertibility conditions for the ARIMA models is a strong indication that this conjecture is likely to be true. In this case, the previous comments on the generality of the steady DLM carry over to these more general models.

In Chapter 10, the Cramér-Rao bound is considered for estimators of the state vector $\underline{\theta}_t$ of the DLM, but now assuming multivariate normal observations and no system error. Observability again plays an important role. It is shown that the information matrix is invertible, and there is a unique estimator which achieves the Cramér-Rao bound if and only if the DLM is observable. Furthermore, it is shown that if the initial estimator $\hat{\underline{\theta}}_o$ for the Kalman filter has covariance matrix equal to the Cramér-Rao lower bound, then all the estimators of $\underline{\theta}_t$ $(t > 0)$ produced by the Kalman filter also achieve the Cramér-Rao bound. This problem is discussed in varying degrees of generality.

REFERENCES

AKAIKE, H. (1974a). Markovian representation of stochastic processes, and its application to the analysis of autoregressive-moving average processes. Ann. Inst. Statist. Math., 26, 363-383.

AKAIKE, H. (1974b). A new look at statistical model identification. I.E.E.E. Trans. on Automatic Control. AC-19, 716-723.

AKAIKE, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. Biometrika, 66, 237-242.

AOKI, M. (1967). Optimisation of stochastic systems. New York: Academic Press.

ANDERSON, O.D. (1975b). The recursive nature of the stationarity and invertibility restraints on the parameters of mixed autoregressive-moving average processes. Biometrika, 62, 704-706.

ANDERSON, O.D. (1977). A further note on the stationarity and invertibility restraints on the parameters of mixed autoregressive-moving average processes. Statist. Hefte, 18, 49-52.

ANDERSON, T.W. (1975a). Maximum likelihood estimation of parameters of autoregressive processes with moving average residuals and other covariance matrices with linear structure. Ann. Statist., 3, 1283-1304.

ANGELL, I.O. and GODOLPHIN, E.J. (1978). Implementation of the direct representation for the maximum likelihood estimator of a Gaussian moving average process. J. Statist. Comput. Simul., 8, 145-160.

ASTROM, K. (1970). Introduction to Stochastic Control Theory. New York: Academic Press.

BARTLETT, M.S. (1946). On the theoretical specification and sampling properties of autocorrelated time series. J.R. Statist. Soc., B, 8, 27-41.

BARTLETT, M.S. and DIANANDA, P.H. (1950). Extensions of Quenouille's test for autoregressive schemes. J.R. Statist. Soc., B, 12, 108-115.

BHANSALI, R.J. (1980). Autoregressive and window estimates of the inverse correlation function. Biometrika, 67, 551-567.

BHANSALI, R.J. and DOWNHAM, D.Y. (1977). Some properties of the order of an autoregressive model selected by a generalisation of Akaike's FPE criterion. Biometrika, 64, 547-551.

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

BOX, G.E.P. and PIERCE, D.A. (1970). Distribution of autocorrelations in autoregressive integrated moving average time series models. J. Amer. Statist. Ass., 65, 1509-1526.

BROWN, R.G. (1959). Statistical Forecasting for Inventory Control. New York: McGraw-Hill.

BROWN, R.G. (1962). Smoothing, Forecasting and Prediction of Discrete Time Series. N.J.: Prentice-Hall, Inc.

BROWN R.G. and MEYER, R.F. (1961). The fundamental theorem of exponential smoothing. Oper. Res., 9, 673-685.

CAMPBELL, M.J. and WALKER, A.M. (1976). A survey of statistical work on the Mackenzie River Series of Annual Canadian Lynx trappings for the years 1821-1934, and a new analysis. J.R. Statist. Soc., A, 140, 411-431.

CHATFIELD, C. and PROTHERO, D.L. (1973). Box-Jenkins seasonal forecasting:problems in a case study (with Discussion). J.R. Statist. Soc., A, 136, 295-336.

COHN, A. (1922). Über die Anzahl der Wurzeln einer Algebraischen Gleichung in einem Kreise. Math. 2. Vol. 14, 110-148.

DAVIES, N. and NEWBOLD, P. (1979). Some power studies of a portmanteau test of time series model specification. Biometrika, 66, 153-155.

D'ESOPO, D.A. (1961). A note on forecasting by the exponential smoothing operator. Oper. Res., 9, 686-687.

DURBIN, J. (1959). Efficient estimation of parameters in moving average models. Biometrika, 46, 306-316.

DURBIN, J. (1960). The fitting of time series models. Rev. Inst. Int. Statist., 28, 233-244.

EYKHOFF, P. (1974). System Indentification: Parameter and State Estimation. Wiley.

FELLER, W. (1968). An Introduction to Probability Theory and its Applications. (Third Edition) Wiley.

GANTMACHER, F.R. (1959). Theory of matrices, Vol. II. New York: Chelsea.

GELB, A. (Editor)(1974). Applied Optimal Estimation M.I.T. Press.

GODOLPHIN, E.J. (1975). A direct basic form for predictors of autoregressive integrated moving average processes. Biometrika, 62, 483-496.

GODOLPHIN, E.J. (1976a). On the Cramer-Wold factorization. Biometrika, 63, 367-80.

GODOLPHIN, E.J. (1976b). Comment on a paper by Harrison and Stevens. J.R. Statist. Soc., B, 38, 238-239.

GODOLPHIN, E.J. (1977). A direct representation for the maximum likelihood estimator of a Gaussian moving average process. Biometrika, 64, 375-384.

GODOLPHIN, E.J. (1978). Modified maximum likelihood estimation of Gaussian moving averages using a pseudo-quadratic convergence criterion. Biometrika, 65, 375-384.

GODOLPHIN, E.J. (1980a). A method for testing the order of an autoregressive moving average process. Biometrika, 67, 699-703.

GODOLPHIN, E.J. (1980b). Estimation of Gaussian linear models. Cahiers du Cero, 243-254.

GODOLPHIN, E.J. and HARRISON, P.J. (1973). The prediction of polynomial processes. University of Warwick: Statistics Report.

GODOLPHIN, E.J. and HARRISON, P.J. (1975). Equivalence theorems for polynomial projecting predictors. J.R. Statist. Soc., B, 37, 205-215.

GODOLPHIN, E.J. and STONE, J.M. (1980). On the structural representation for polynomial projecting predictor models based on the Kalman filter. J.R. Statist. Soc., B, 42, 35-46.

GOLDBERGER, A.S. (1964). Econometric Theory. Wiley.

HANNAN, E.J. (1969). The estimation of mixed moving average autoregressive systems. Biometrika, 56, 579-593.

HARRISON, P.J. (1967). Exponential smoothing and short-term sales forecasting. Man. Sci., 13, 321-342.

HARRISON, P.J. and STEVENS, C.F. (1971). A Bayesian approach to short-term forecasting. Oper. Res. Quart., 22, 341-362.

HARRISON, P.J. and STEVENS, C.F. (1975). Bayesian forecasting in action: case studies. University of Warwick: Statistics Research Report.

HARRISON, P.J. and STEVENS, C.F. (1976). Bayesian Forecasting (with Discussion). J.R. Statist. Soc., B, 38, 205-247.

HOLT, C.C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. Carnegie Institute of Technology: ONR Memorandum 52.

JACOBS, O.L.R. (1974). Introduction to Control Theory. Oxford: Clarendon Press.

JAZWINSKI, A.H. (1970). Stochastic Processes and Filtering Theory. New York: Academic Press.

JOSHI, V.M. (1976). On the attainment of the Cramér-Rao lower bound. Ann. Statist., 4, 998-1002.

JURY, E.I. (1964). Theory and Application of the z-transform method. Wiley.

KALMAN, R.E. (1960). A new approach to linear filtering and prediction problems. Trans. A.S.M.E. Ser. D.J. Basic. Eng. 82, 35-45.

KALMAN, R.E. (1961). On the general theory of control systems. Proc. 1st. Int. Conference on Automatic Control, Moscow, 1960. Butterworths, London, 1961, Vol. 1, 481-492.

KALMAN, R.E. (1963a) New Methods in Weiner Filtering Theory. Proc. 1st. Symp. on Eng. Applications of Random Function Theory and Probability Theory. (J.L. Bogdanoff and F. Kozin, Eds.) Wiley.

KALMAN, R.E. (1963b). Mathematical descriptions of linear dynamical systems. Jour. SIAM. Control A, 1, 152-192.

KALMAN, R.E. and BUCY, R.C. (1961). New results in linear filtering and prediction theory. Jour. Basic. Eng. (ASME translation), 83D, 95-108.

KEY, P.B. and GODOLPHIN, E.J. (1981). On the Bayesian steady forecasting model. J.R. Statist. Soc., B, 43 , 92-96.

KUSHNER, H. (1971). Introduction to Stochastic Control. Holt, Rinehart and Winston, Inc.

LINDORFF, D.P. (1965). Theory of Sampled-Data Control Systems. Wiley.

LJUNG, G.M. and BOX, G.E.P. (1978). On a measure of lack of fit in time series models. Biometrika, 65, 297-303.

LOMNICKI, Z.A. and ZAREMBA, S.K. (1957). On the estimation of autocorrelation in time series. Ann. Math. Statist., 28, 140-158.

MANN, H.B. and WALD, A. (1943). On the statistical treatment of linear stochastic difference equations. Econometrica, 11, 173-220.

MOON, J.R. (1977). Random and systematic errors in some common expanding memory and fading memory filters. Internal Ferranti Report SPAT 279.

MORRISON, N. (1969). Introduction to Sequential Smoothing and Prediction. McGraw-Hill, Inc.

MUTH, J.F. (1960). Optimal properties of exponentially weighted forecasts. J. Amer. Statist. Ass., 55, 299-306.

NEWBOLD, P. and GRANGER, C.W.J. (1974). Experience with forecasting univariate time series and the combination of forecasts (with Discussion). J.R. Statist. Soc., A, 137, 131-164.

OPPENHEIM, A.V. and SCHAFER, R.W. (1975). Digital Signal Processing. Prentice Hall, Inc.

PAGANO, M. (1973). When is an autoregressive scheme stationary? Comm. Statist., 1, 533-544.

PHAM-DINH, T. (1979). The estimation of parameters for autoregressive moving average models from sample autocovariances. Biometrika, 66, 555-560.

PRICE, C.P. (1974). An introduction to α-β tracking. Internal Ferranti Report SPAT 134.

PRIESTLEY, M.B (1980). A general approach to non-linear time-series analysis. Cahiers du Cero. 285-307.

PROTHERO, D.L. and WALLIS, K.F. (1976). Modelling macro-economic time series (with Discussion). J.R. Statist. Soc., A, 139, 468-500.

QUENOUILLE, M.H. (1947a). Notes on the calculation of autocorrelations of linear autoregressive schemes. Biometrika, 34, 365-367.

QUENOUILLE, M.H. (1947b). A large sample test for the goodness of fit of autoregressive schemes. J.R. Statist. Soc., A, 110, 123-129.

RAO, C.R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. Bull. Calcutta. Math. Soc., 37, 81-91.

RAO, C.R. (1973). Linear Statistical Inference and its Applications. (Second Edition). Wiley.

RAY, W.D. and WYLD, C. (1965). Polynomial projecting predictors: properties of multi-term predictors/ controllers in non-stationary time series. J.R. Statist. Soc. B, 27, 144-158.

ROUTH, E.J. (1905). The advanced part of a treatise on the Dynamics of a System of Rigid Bodies. Macmillan and Co., London.

SCHUR, I. (1917). Über Potenzreichen, die im Innen des Einheitskreises beschränkt sind. Journal für Mathematik, 147, 202-232.

SHAMAN, P. (1976). Approximations for stationary covariance matrices and their inverses, with application to ARMA models. Ann. Statist., 4, 292-301.

SHERMAN, S. (1955). A theorem on convex sets with applications. Ann. Math. Stat., 26, 763-767.

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

SMITH, J.Q. (1979). A generalisation of the Bayesian steady forecasting model. J.R. Statist. Soc., B, 41, 375-387.

SORENSON, H.W. (1966). Kalman Filtering Techniques. Advances in Control Systems, Vol. 3. (C.T. Leondes, Ed.). New York: Academic Press.

TONG, H. (1977). Some comments on the Canadian Lynx data. J.R. Statist. Soc., A, 140, 432-436, 448-468.

TONG, H. and LIM, K.S. (1980). Threshold autoregression, limit cycles and cyclical data. J.R. Statist. Soc., B, 42, 777-858.

TUNNICLIFFE-WILSON, G. (1969). Factorisation of the covariance generating function of a pure moving average process. S.I.A.M.J. Numer. Anal. 6, 1-7.

WALKER, A.M. (1952). Some properties of the asymptotic power functions of goodness of fit tests for linear autoregressive schemes. J.R. Statist. Soc., B, 14, 117-134.

WALKER, A.M. (1961). Large sample estimation of parameters for moving average models. Biometrika, 48, 343-357.

WALKER, A.M. (1962). Large sample estimation of parameters for autoregressive processes with moving average residuals. Biometrika, 49, 117-132.

WHITTLE, P. (1951). Hypothesis testing in Time Series Analysis. Uppsala: Almquist and Wiksell.

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

WHITTLE, P. (1953). Estimation and information in stationary time series. Ark. Mat. Fys. Astr. 2,423-434.

WHITTLE, P. (1954). Some recent contributions to the theory of stationary processes. Appendix 2 in Wold (1954).

WHITTLE, P. (1963). Prediction and Regulation by Linear Least-Squares Methods. Princeton: Van Nostrand.

WHITTLE, P. (1969). A view of stochastic control theory. J.R. Statist. Soc., A, 132, 320-334.

WINTERS, P.R. (1960). Forecasting sales by exponentially weighted moving averages. Man. Sci., 6, 324-342.

WISE, J. (1956). Stationarity conditions for stochastic processes of the autoregressive and moving average type. Biometrika, 43, 215-219.

WISHART, D.M.G. (1969). A survey of control theory. J.R. Statist. Soc., A, 132, 293-319.

WOLD, H.O.A. (1949). A large sample test for moving averages. J.R. Statist. Soc., B, 11, 297-305.

WOLD, H.O.A. (1954). A study in the Analysis of Stationary Time Series. (Second Editon). Stockholm Academic Press.

YAGLOM, A.M. (1955). The correlation theory of processes whose nth difference constitute a stationary process. Matem. Sb. 37, 141-196.

YAGLOM, A.M. (1962). An Introduction to the Theory of Stationary Random Functions. Translated by R.A. Silverman. Englewood Cliffs, N.J. Prentice-Hall Inc.

ZACKS, S. (1971). The Theory of Statistical Inference. Wiley.