B A Y E S I A N F O R E C A S T I N G

W I T H

S T A T E S P A C E M O D E L S

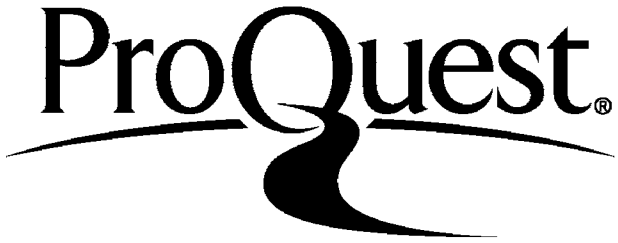
by

Peter Bernard Key

Royal Holloway
and
Bedford New College

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

November 1985



ABSTRACT

This thesis explores the use of State-Space models in Time Series Analysis and Forecasting, with particular reference to the Dynamic Linear Model (DLM) introduced by Harrison and Stevens. Concepts from Control Theory are employed, especially those of observability, controllability and filtering, together with Bayesian inference and classical forecasting methodology.

First, properties of state-space models which depart from the usual Gaussian assumptions are examined, and the predictive consequences of such models are developed. These models can lead to new phenomena, for example it is shown that for a wide class of models which have a suitably defined steady evolution the usual properties of classical steady models (such as exponentially weighted moving averages) do not apply.

Secondly, by considering the forecast functions, equivalence theorems are proved for DLMs in the steady state and stationary Box-Jenkins models. These theorems are then extended to include both time-varying and non-stationary models thus establishing a very general predictor equivalence. However it is shown that intuitively appealing DLMs which have diagonal covariance matrices are restricted by only covering part of the equivalent stability / invertibility region, and examples are given to illustrate these points.

Thirdly, some problems of inference involving state-space models are looked at, and new approaches outlined. A class of collapsing procedures based upon a distance measure between posterior components is introduced. This allows the use of non-normal errors or Harrison-Stevens Class II models by condensing the normal-mixture posterior distribution to prevent an explosion of information with time, and avoids some of the problems of the Harrison-Stevens solution.

Finally, some examples are given to illustrate the way in which some of these models and collapsing procedures might be used in practice.

To my parents.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

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 in Chapters 5-10 is believed to be original.

The support of S.E.R.C. finance is gratefully acknowledged. In addition, my thanks must go to various members of the Statistics and Mathematics Departments of Royal Holloway College for helpful discussions - in particular Janice Stone and David Yates, to Jim Smith of University College for awakening my interest in Bayesian Forecasting, and to my wife Fiona for patience and assistance with the typing. Lastly, I would especially like to thank my supervisor Ed Godolphin for guidance and encouragement.

CHAPTER 1

INTRODUCTION

Time plays an important rôle in many situations which involve data, for example much economic data consists of quantities or statistics which vary with time. In the context of experimentation although strictly speaking each observation is made at a unique time, in some instances time is a necessary ingredient of the analysis. This means that the ordering of the data matters, and it is this dependence between data items that complicates the analysis.

Throughout this thesis we shall be concerned with situations such as these, where we have a set of observations or data indexed by t, $\{y_t\}$ say. We shall only look at the case where time is discrete, with observations made at equally spaced intervals, that is a Time Series. Unless otherwise stated we shall further assume that the observations y_t are univariate, and let the indexing parameter t range over a subset of the integers.

A time series might be analysed for a variety of reasons, for example:

(1) To generate information about the underlying process from which the observations came; this might lead on to questions of model determination or hypothesis testing.

(2) In a control situation, where the observations are arriving in 'Real Time' and we wish to alter certain physical parameters to achieve a specified objective.

(3) To enable us to make statements or forecasts about future y_t.

This list is neither exhaustive nor indeed are its items mutually exclusive. However, there is a difference in emphasis between them and different criteria are used to judge model performance.

We shall mainly be concerned with the last objective given above, namely (3). That is we have observations $\{y_t\}$ up to some point t_o, and we wish to make statements about y_t for $t>t_o$. In this case our success is measured by the quality of our forecasts, namely how close our forecasts or predictions are to the subsequently observed values, where closeness is suitably defined.

To be able to analyse time series statistically we require an appropriate mathematical framework, which is the class of discrete stochastic processes. A discrete stochastic process is a family of random variables $\{y_t\}$ indexed by t where t varies over the integers. Strictly speaking we should distinguish in notation between the underlying mathematical model and a particular observed time series or realisation of the stochastic process. Indeed it is important to remember that any model we build is an approximation, which we use to draw inferences using the actual data.

It is necessary to impose additional structure if we are to be able to make meaningful statements about a particular time series, and now two cases arise. The first is where there is an underlying model which presents itself from physical or theoretical considerations, or perhaps a model specified up to a certain point. For example, if the time series is the observed position of a space-craft at equally spaced intervals of time, then successive positions

are related by the equations of motion. However, in a more comprehensive model uncertain factors - such as wind-strength if in the atmosphere, errors of measurement and other unmeasured forces - can be represented probabilistically. To represent this, the deterministic equations of motion are incorporated in a probabilistic model.

The second case is where we have no firm idea of the nature of the underlying model. This might be because one is not known or because there are too many factors to take account of. It is these type of time series which are most commonly studied in Statistics and indeed in Economics. One is left with little more than the data from which to deduce the structure of the series.

In an attempt to make this thesis largely self-contained, Chapters 2-4 are of an introductory nature, describing the background against which Chapters 5-9 are set.

Chapter 2 gives a resumé of some of the main ideas that are used in classical statistical Time Series analysis, which is linked with objective (1) above. The first step in developing a tractable class of time series models is to introduce the concept of stationarity, and in particular second-order stationarity, which requires that the first and second order properties of the time series are independent of time.

It is then possible to use the so-called 'Time Domain' or 'Frequency Domain' approaches. The first deals with properties of the time series per se, whilst the second in essence performs an harmonic analysis, which gained impetus from the widespread use of Fourier analysis in the engineering disciplines. This latter approach leads to spectral

decomposition and the identification of frequencies, for example as described by Hannan (1960).

The simplest additional structure to impose is linearity, which is strongly linked to the Gaussian distribution. The general ARMA (auto-regressive moving average) models express a linear combination of past and present observations as a linear combination of past and present 'noise' variables. Non-stationary models can be generated by suitably differencing the series to give a stationary ARMA model, creating the ARIMA models - I for integrated. The popularity and widespread use of such models owes much to the work of Box-Jenkins (1970) and Jenkins (1979).

As we have mentioned, another use of time series is in a control environment. In the example of a space-craft given above, not only do we wish to be able to predict the position of the space-craft at future instants of time, but also we want to be able to alter its position by applying the appropriate thrusts to achieve a desired objective, such as landing on the moon.

The illustration of a space-craft is no accident: 'classical' control theory is the analysis of differential equations or difference equations, often by transform methods, which received a boost from military applications in the Second World War. But modern control theory which uses a state-space approach, benefitted from the enormous amount of reseach undertaken during the American Space programme. In the state-space approach the system is described by a state vector, each of whose components are physical quantities of interest, such as the position and momentum co-ordinates of the vehicle. However normally

these are not all directly available, instead observations contain information about part of the state vector. This relationship is specified, as is the evolution of the state vector in time, and the object is to estimate the state vector, or control it by applying suitable inputs.

Chapter 3 introduces some of the key concepts in control theory, especially those of the pioneering work of Kalman (1963) on controllability and observability, and various types of stability are mentioned and linked together The idea of equivalence between different descriptions is introduced.

When the error terms in linear state-space models are Gaussian, the Kalman filter provides optimal estimates of the state vector. This is one of the concepts which has been part of a healthy cross-fertilisation between the disciplines of Engineering (in particular control theory) and Statistics.

Harrison and Stevens in particular (1971, 1975, 1976) have shown that it is possible to usefully use concepts from control theory in Time Series analysis. At the heart of their theory is the Dynamic Linear Model (DLM), which is a linear state-space model with additive Gaussian noise, upon which the Kalman filter is used. These models, together with extensions developed by Harrison and Stevens, are discussed in Chapter 4, which also includes a brief introduction to the Bayesian paridigm.

In many branches of statistics, distributions are assumed to be Gaussian, or 'normal' and indeed it is often very difficult work outside this framework. Chapter 5 examines the possibility of using non-normal state-space

models in the context of Bayesian analysis. Unless the error distributions are stable, it is not easy to introduce non-normality into the state-space description and obtain tractable results. Smith (1979) used a slightly indirect method of extending the simplest DLM - the steady model - to include non-normality. The implications for the predictors, or the predictive distributions, of using such a steady evolution are examined. In particular, the exponential family is used, and it is shown that when a steady evolution is used with members of this family, that many of the familiar properties of the steady model fail to hold. For example, the predictors at a particular time for different lead times need not be identical, as they are for the Normal Steady model. Some of these results have already been discussed in Key and Godolphin (1981).

Certain types of equivalence between state-space and ARIMA models have been discussed before, such as in Godolphin and Harrison (1975). Chapter 6 looks at predictor equivalence between DLMs and ARIMA models, that is models which yield the same predictors for all lead times. First the role of observability is established in this context - essentially one only has to consider the observable subsystem of any DLM. Then constant DLMs which are in the steady state (for which conditions are given in Chapter 4) are shown to be equivalent to ARIMA models with constant coefficients. These results encompass both stationary and non-stationary models. It is then further shown that DLMs which are time-varying (or constant ones not yet in the steady-state) are equivalent to ARIMA models with varying parameters.

An alternative way to demonstrate equivalence is to use the covariance properties of the models, and subject to mild restrictions, the equivalence classes are isomorphic to those generated by looking at predictors. A slightly surprising result is that in many cases unless the system matrix of the DLM is singular, the parameters in the equivalent ARIMA model can be severely restricted.

Chapter 7 examines the consequences of Chapter 6 as they apply to practical DLMs. Theoretical results are proved which describe the invertibility regions that are mapped out by DLMs having constant system and observation matrices, but whose covariance matrices vary. It is shown that practical DLMs, which have diagonal covariance matrices, can only cover part of the invertibility region of the equivalent ARIMA models, except in the case of first order models. A slightly different way of writing some useful DLMs is given, which calls for a slight ammendment to the Kalman filter. Finally, a summary is provided of some frequently used ARIMA models and their DLM equivalents.

The Bayesian forecasting methodology of Harrison-Stevens (1976) not only included DLMs but also introduced the idea of multi-process models. In other words, one can assume that different DLMs are operative at different times, with a transition matrix describing the evolution between them. The problem with this generality is in 'collapsing' the ever increasing number of normal components of the posterior to a smaller number. A simple approach is given in Harrison-Stevens (ibid). However, Chapter 8 describes a new class of collapsing procedures, based upon a clustering approach with an appropriate metric. In particular, the

Hellinger distance is examined and certain desirable properties of the collapsing procedure proved.

Chapter 9 illustrates how some of the preceeding ideas might be applied in practice by looking at two real time series. The augmented steady model is used for simplicity, together with the collapsing procedure of Chapter 8.

CHAPTER 2

TIME SERIES

2.1 Introduction

At the present time there is much interest in time series analysis, partly because it finds applications in very diverse fields and partly because of the problems involved, even in simple models. We take a time series to mean a series of observations taken at discrete points in time, and only consider univariate observations. The corresponding probability model is a family of random variables, $\{y_t\}$, where without loss of generality the index t ranges over the integers.

The simplest models are linear and much current practice using these has been influenced by the seminal work of Box and Jenkins (1976). This theoretical work is complemented by Jenkins (1979) who illustrates the use of such models in practical situations. There is a great deal of literature on all aspects of time series, and useful reviews are provided by Chatfield (1977) and Cox (1981).

First we consider stationary series, which form the backbone of the subject, and then describe some models, how to forecast with them and discuss the problems of inference.

2.2 Stationary Time Series

A time series $\{y_t\}$ is said to be second-order (or

weakly) stationary if $E(y_t) = \mu$ is a constant for all t and $\text{cov}(y_t, y_{t+h})$ is a function of h alone. We can therefore describe the second-order properties by the autocovariance function

$$\gamma_h = \text{cov}(y_t, y_{t+h})$$

where h is an integer, or by the autocorrelation function

$$\rho_h = \gamma_h/\gamma_0 = \text{cor}(y_t, y_{t+h}).$$

It follows that $\gamma_h = \gamma_{-h}$, $\rho_0 = 1$, $\rho_h = \rho_{-h}$.

Other descriptors are possible, for instance the spectral distribution function $G(\omega)$ and density function $g(\omega)$ come from a harmonic decomposition of y_t given by

$$\gamma_t = \int_{-\pi}^{\pi} e^{i\omega t}\, dG(\omega)$$

$$g(\omega) = G'(\omega) = \tfrac{1}{2}\pi \sum_{-\infty}^{\infty} e^{i\omega s}\, \gamma_s \qquad -\pi \leq \omega \leq \pi.$$

This forms part of the Frequency Domain approach to time series as expounded by Hannan (1960) say. We shall concentrate on the so-called Time Domain approach.

For any discrete stationary process $\{y_t\}$ the Wold decomposition theorem states that the process can be represented as the sum of two mutually uncorrelated processes $\{x_t\}$ and $\{z_t\}$, $y_t = x_t + z_t$, where

(i) x_t is deterministic

(ii) z_t is a purely non-deterministic moving average

$$z_t = \sum_{j=0}^{\infty} b_j \varepsilon_{t-j}, \quad b_0 = 1$$

where $\sum_{j=0}^{\infty} b_j^2 < \infty$ and ε_t is a sequence of uncorrelated random variables of zero mean and finite variance. $\{x_t\}$ and $\{z_t\}$ are uniquely specified and either may be absent.

2.3 Models for Time Series

Early time series models were often of the form

$$y_t = m_t + s_t + z_t$$

where m_t is a trend term, depicting a smooth long term movement, s_t a seasonal term - periodic oscillations of known frequency - and z_t an error term, usually of zero mean. Multiplicative models and mixed additive/multiplicative models can similarly be defined in these terms. The analysis might then proceed by fitting some form of polynomial to m_t or s_t, the former being the basis of the much used moving average trend removal techniques (Kendall 1976). There are problems concerning this more or less empirically based approach, not least concerning the reasonableness of the assumptions. Modern classical time series analysis has tended towards a more theoretical model based approach.

The widely used ARMA (p,q) models, auto-regressive moving average models of order p,q are defined by

$$y_t + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} \qquad (2.1)$$

with ε_t pure noise, that is random variables of zero mean and common variance σ^2. In fact the ARMA (p,q) is usually taken to mean the above, (2.1), with the conditions $|\lambda_i|<1$, $|\mu_i|<1$ where λ_i and μ_i are the roots of the polynomials

$$\alpha(z) = 1 + \alpha_1 z + \ldots + \alpha_p z^p = \prod_{i=1}^{p} (1-\lambda_i z)$$

and

$$\beta(z) = 1 + \beta_1 z + \ldots + \beta_q z^q = \prod_{i=1}^{q} (1-\mu_i z).$$

$|\lambda_i|<1$ ensures that the model in stationary, so that the

non-deterministic time series $\{y_t\}$ has an infinite moving average representation $\sum_{j=0}^{\infty} b_j \varepsilon_{t-j}$; (2.2)

the second condition is the invertibility relation which ensures that the model has an infinite autoregressive representation

$$\varepsilon_t = \sum_{j=0}^{\infty} a_j x_{t-j} \qquad \text{with } a_0 = 1.$$

Each y_{t-j} in (2.1) can be replaced by $y_{t-j} - \mu$ so that we can assume without loss of generality the model has zero mean.

The autocovariance function can be found from (2.2) by multiplying by y_{t+h} and taking expectations, giving

$$\gamma_h = \sigma^2 \sum_{i=0}^{\infty} b_i b_{i+h} \qquad (2.3)$$

where $\sigma^2 = \text{var}(\varepsilon_t)$. Defining the generating functions

$$A(z) = \sum_{j=0}^{\infty} a_j z^j, \quad B(z) = \sum_{j=0}^{\infty} b_j z^j, \quad \Gamma(z) = \sum_{k=-\infty}^{\infty} \gamma_k z^k$$

then for an ARMA (p,q) process $A(z) = \dfrac{1 + \alpha_1 z + \ldots + \alpha_p z^p}{1 + \beta_1 z + \ldots + \beta_q z^q}$

and $B(z)=1/A(z)$ enabling the coefficients b_i to be found. Alternatively the well known relation $\Gamma(z)=\sigma^2/\{A(z)A(z^{-1})\}$ can be used to determine the autocovariances.

Example 2.1

Consider the ARMA(1,1) model

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

Then $B(z) = (1+\beta z)/(1+\alpha z)$ and equating powers of z gives

$$b_0=1, \quad b_i = (-\alpha)^{i-1}(\beta-\alpha), \quad i=1,2 \ldots .$$

so that γ_h is given by (2.3). Using generating functions

$$\Gamma(z) = \sigma^2 \frac{1+\beta z}{1+\alpha z}\frac{1+\beta z^{-1}}{1+\alpha z^{-1}}.$$

If we write $\frac{\Gamma(z)}{\sigma^2} = A_0 + \frac{A_1 z}{1+\alpha z} + \frac{A_1 z^{-1}}{1+\alpha z^{-1}}$

then equating powers of z gives

$$A_0 = \frac{1 + \beta^2 - 2\alpha\beta,}{1 - \alpha^2} \qquad A_1 = \frac{(1 - \alpha\beta)(\beta - \alpha)}{1 - \alpha^2}$$

but $\Gamma(z) = \gamma_0 + \sum_{k=0}^{\infty} \gamma_k z^k + \sum_{k=0}^{\infty} \gamma_k z^{-k}$ so that

$$\gamma_0 = A_0\sigma^2, \qquad \sum_{k=0}^{\infty} \gamma_k z^k = \sigma^2 \frac{A_1 z}{1+\alpha_1 z} .$$

Thus equating coefficients of z^k

$$\gamma_0 = \sigma^2\frac{(1+\beta^2-2\alpha\beta)}{1-\alpha^2} , \qquad \gamma_1 = \sigma^2\frac{(1-\alpha\beta)(\beta-\alpha)}{1-\alpha^2}$$

and since $\gamma_k + \alpha\gamma_{k-1} = 0$, $\gamma_k = (-\alpha)^{k-1}\gamma_1 , \quad k \geq 1.$

This approach is a specific example of Quenouilles algorithm for obtaining the autocovariances using generating functions.

If we denote the backward shift operator by T, so that $Ty_t = y_{t-1}$, $T^k y_t = y_{t-k}$ then the ARMA(p,q) model can be written as

$$\alpha(T)y_t = \beta(T)\varepsilon_t .$$

These models can only represent stationary time series. A class of models for non-stationary models can be obtained by first differencing the series d times to give the transformed series x_t

$$x_t = (1 - T)^d y_t$$

and then fitting an ARMA(p,q) model to x_t. These are

called ARIMA (p,d,q) models, and can be written

$$\alpha(T)(1-T)^{d}y_t = \beta(T)\varepsilon_t.$$

In practical situations d is usually chosen to be small, for example a model which we shall meet later is the ARIMA (0,1,1) or IMA (1,1) model

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

with single differencing. In general single differencing describes a process whose level is continually updated and double differencing introduces a slope term which is also updated.

Seasonal models can be described by similar models using a difference operator of the required periodicity. For example if s is the period then an appropriate model is

$$\alpha(T^s)(1 - T^s)y_t = \beta(T^s)\varepsilon_t$$

and more generally multiplicative $(p,d,q) \times (P,D,Q)s$ models

$$\alpha_p(T)\alpha_P(T^s)(1-T)^{d}(1-T^s)^{D}y_t = \beta_q(t)\beta_Q(T^s)\varepsilon_t$$

where α_p is a polynomial of degree p and so on.

2.4 Forecasting

In many situations we have observations up to time t, $y_t, y_{t-1}, \ldots$, which from now on we write as y^t, and we wish to predict y_{t+m} for m greater than or equal to one. The predictor $y_t(m)$ of y_{t+m} that minimises the mean square error

$$\sigma_t^2(m) = E\{y_t(m) - y_{t+m}\}^2 \qquad (2.4)$$

is the conditional expectation

$$y_t(m) = E(y_{t+m}|y^t). \tag{2.5}$$

In general a lot of information about the time series is required to be known before this quantity can be calculated, however it is relatively easy to calculate for linear models with a simple error structure.

Less information is required if we restrict $y_t(m)$ to a linear function of the past data

$$y_t(m) = \sum_{j=0}^{\infty} d_j(m) y_{t-j} \tag{2.6}$$

and seek to minimise the mean square error (2.4). We then only need the first and second order properties of the process, although in practice even these will only be known approximately from the observations. The resulting predictor will then be best if and only if the conditional expectation (2.5) is linear.

From the Wold decomposition theorem a purely non-deterministic stationary series can be written as

$$y_t = \sum_{j=0}^{\infty} b_j \varepsilon_{t-j}. \tag{2.7}$$

The minimum mean square error (MMSE) linear predictor of $y_t(m)$ is then given by

$$y_t(m) = \sum_{j=0}^{\infty} b_{j+m} \varepsilon_{t-j} \tag{2.8}$$

with mean square error $\sigma_t^2(m) = \sigma^2 \sum_{j=0}^{m-1} b_j^2$. (2.9)

This gives the forecasts in terms of the random errors rather than the observations which is what we require. If we let

$$D_m(z) = \sum_{j=0}^{\infty} d_j(m) z^j$$

then the forecast weights $d_j(m)$ can be calculated from the generating function relation

$$D_m(z) = \frac{\sum_{j=0}^{\infty} b_{j+m} z^j}{B(z)} \qquad (2.10)$$

where as in §2.3 $B(z) = \sum_{0}^{\infty} b_j z^j$. The $d_j(m)$ can be calculated from (2.10) and substituted into (2.6) to give the forecasts provided that (2.10) can be expressed as a power series. This condition holds true for ARMA models proved that they are invertible.

Three consequences of the above are

(i) $E(y_{t+m}|y^t) = \sum_{j=0}^{\infty} b_j E(\varepsilon_{t+m-j}|y^t)$ using the Wold decomposition which since

$$E(\varepsilon_{t+m-j}|y^t) = 0 \quad j = 0, 1 \ldots m-1$$
$$= \varepsilon_t \text{ otherwise}$$

is identical to (2.8). The predictor is therefore the MMSE predictor and is obtained form the Wold decomposition by setting all future random disturbances to zero, their expectation.

(ii) The mean square predictor error $\sigma_t^2(m)$ increases with m to the limit $\sigma^2 \sum b_j^2 = \gamma_0$. The smallest error is for lead time 1, that is m = 1 and $\sigma_1^2 = \sigma^2$.

(iii) The sequence ε_{t+1}, ε_{t+2} ... are the one step ahead predictor errors.

Example 2.2

For the ARMA (1,1) process of example 2.1

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

the coefficients in the Wold decomposition are $b_i = (-\alpha)^{i-1}(\beta-\alpha)$ for $i>1$, and $B(z) = (1+\beta z)/(1+\alpha z)$. Substituting in (2.10)

$$D_m(z) = \left(\frac{1+\alpha z}{1+\beta z}\right) \sum_{j=0}^{\infty} (-\alpha)^{m-1+j}(\beta-\alpha)z^j$$

$$= (-\alpha)^{m-1}\left(\frac{\beta-\alpha}{1+\beta z}\right)$$

equating powers of z to obtain $d_j(m)$ and substituting in (2.6) gives

$$y_t(m) = (-\alpha)^{m-1}(\beta-\alpha) \sum_{j=0}^{\infty} (-\beta)^j y_{t-j} \qquad (2.11)$$

The mean square prediction errors of (2.4) are given from (2.9)

$$\sigma_t^2(m) = \sigma^2\left[1 + \frac{(\alpha-\beta)^2\{1-\alpha^{2(m-1)}\}}{1-\alpha^2}\right].$$

Alternatively, at time $t+1$

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

taking conditional expectations at time t

$$y_t(1) = -\alpha y_t + \beta\varepsilon_t \qquad (2.12)$$

and similarly at lead time m

$$y_t(m) = -\alpha y_t(m-1). \qquad (2.13)$$

Expressing ε_t in terms of the past data y^t, which we can do since the model is invertible, and substituting in (2.12) and (2.13) gives (2.11).

At time $t+1$, (2.12) is

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

showing how the forecasts can be updated when a new observation arrives.

Although we have only dealt with stationary models so far, the second method of the above example can be

used to forecast non-stationary models as follows. First define $y_t(m)$ and $\sigma_t^2(m)$ as in (2.4) and (2.5), so that $\sigma_t^2(m) = \text{var}\,(y_{t+m}|y^t)$. Now the general class of ARIMA models, including seasonal variants, which were described in §2.3 can all be expressed in difference equation form as

$$\phi(T)y_t = \theta(T)\varepsilon_t \qquad (2.14)$$

where $\theta(z)$ is a polynomial with all its zeros outside the unit circle, ensuring that the model is invertible, and $\phi(z)$ has no zeros inside the unit circle.

Forecasts can then be calculated by using (2.14) at times t+1, t+2 ... and taking conditional expectations at time t using

$$E(y_{t+j}|y^t) = y_t(j) \qquad j = 1, 2, \ldots$$
$$E(y_{t-j}|y^t) = y_{t-j} \qquad j = 0, 1, \ldots$$
$$E(\varepsilon_{t-j}|y^t) = \varepsilon_{t-j} = y_{t-j} - y_{t-j-1}(1) \qquad j = 0, 1, \ldots$$
$$E(\varepsilon_{t+j}|y^t) = 0. \qquad j = 1, 2, \ldots$$

This algorithm produces a difference equation in terms of the observations y_{t-j} and predictors $y_{t-j}(1)$, $y_t(j)$ which can then be solved to give $y_t(m)$ in terms of the past observations.

The prediction errors at lead time m are given in Box and Jenkins (1976, p 128) as

$$e_t(m) \equiv y_{t+m} - y_t(m) = c_0\varepsilon_{t+m} + c_1\varepsilon_{t+m-1} \cdots c_{m-1}\varepsilon_{t+1}.$$

The c_j are the coefficients of the random terms when the observation is expressed as an infinite weighted sum of current and previous shocks

$$y_t = \sum_{j=0}^{\infty} c_j\varepsilon_{t-j}$$

so that under (2.14) the c_j's are obtained from

$$\phi(T)(1 + c_1 T + c_2 T^2_t \ldots) = \theta(T).$$

For stationary models $\phi(z)$ has zeros all outside the unit circle so that the c_j's are identical to the b_j's of (2.7).

It follows that the forecasts are unbiased with prediction variance $\sigma^2_t(m) = \sigma^2(1 + c_1^2 + \ldots + c^2_{m-1})$

$$= \sigma^2_t(m-1) + c^2_{m-1}\sigma^2.$$

The coefficients c_j also enable us to up-date the forecasts:

At time t $\quad y_t(m) = y_{t+m} - e_t(m),$

if a new observation arrives

$$y_{t+1}(m-1) = y_{t+m} - e_{t+1}(m-1),$$

but

$$e_t(m) - e_{t+1}(m-1) = c_{m-1}\varepsilon_{t+1} = c_{m-1}e_t(1)$$

and on substituting

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

Example 2.3

Consider the ARIMA (0,1,1) model

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

At time t+1 $\quad y_{t+1} - y_t = \varepsilon_{t-1} + \beta\varepsilon_t$

so taking conditional expectations and using the above algorithm gives

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

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

which has solution

$$y_t(1) = (1+\beta)\sum_{j=0}^{\infty}(-\beta)^j y_{t-j}. \qquad (2.15)$$

For $k \geq 2$, writing down the model at time t + k and taking expectations conditional upon y^t gives

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

which completes the specification of the forecast function. The c_j can be obtained from

$$(1 - T)(1 + c_1 T + c_2 T^2 + \ldots) = (1 + \beta)$$

equating powers of T, $c_i = 1 + \beta \quad i \geq 1$

so that

$$\sigma_t(m) = \sigma^2\{1 + (m - 1)(1 + \beta)^2\}.$$

The above results are identical to those of Example 2.2 with α formally replaced by $\alpha = -1$.

Apart from simple cases, the difference equation produced by the algorithm can be difficult to solve. Godolphin (1975) introduces an updating and component series which enables $y_t(m)$ to be calculated without first calculating the forecasts of $y_{t+1}, \ldots y_{t+m-1}$.

It follows from (2.14) and the forecasting algorithm given above that for $k>q$, where q is the order of the polynomial $\theta(z)$ that

$$\phi(T)y_t(k) = 0 \qquad (2.16)$$

where T operates on k, so that $Ty_t(k) = y_t(k-1)$ and so on. This means that the eventual forecast function is determined solely by $\phi(z)$.

If $\phi(z^{-1})$ has r roots λ_i with multiplicity m_i, so $\sum m_i = p$ then the general solution of (2.16) is

$$y_t(k) = \sum_{i=1}^{r} \sum_{j=0}^{m_i - i} h_{ij} k^j \lambda_i^k$$

where the constants h_{ij} are determined from the first p forecasts, and so depend upon the data and the parameters $\alpha_1 \ldots \alpha_p \beta_1 \ldots \beta_q$.

For example

$$\phi(z) = 1 - z, \quad y_t(k) = h_1, \text{ a constant } k>1$$

$$\phi(z) = (1-z)^2 \quad y_t(k) = h_1 + h_2 k$$

which is the so called linear growth model.

Some methods of forecasting involve choosing the eventual forecast function first and then fitting the coefficients from the data, whereas Box and Jenkins fit the model first, and so get the eventual forecast function as a consequence.

2.5 Parameter Estimation

In the Box-Jenkins approach to time-series, first a model is tentatively identified, then parameters are fitted, diagnostic checks are applied and if inadequacy is shown then the iterative process of identification, estimation and checking is repeated until a suitable model is found. Each of these three stages is a subject in itself with its own extensive literature, which we shall not go into. Apart from the theoretical side, examining real data poses its own problems; for example the famous Lynx data has been analysed over the years by many authors using different models. Even within the same framework practitioners can fit different models to the same series, for example Chatfield and Prothero (1973).

The first step with non-stationary models is to try and transform to a stationary time-series. In the ARIMA (p,d,q) models (and seasonal variants) this is achieved by differencing the series. Box and Jenkins suggest looking at the correlogram, which is the plot of the sample

autocorrelations r_h against h where

$$r_h = c_h / c_0 \tag{2.17}$$

$$c_h = \frac{1}{N-h} \sum_{t=1}^{N-h} y_t y_{t+h} \tag{2.18}$$

and N is the number of observations. There are possible variants of (2.18) such as having the denominator N rather than N - k. These are model-free estimators of autocorrelations ρ_h of §2.2 but are best viewed as estimating the autocorrelation function ρ_h rather than individual values.

The sampling properties of the individual r_h's depend upon all the ρ_h. However for stationary series the estimates are asymptotically unbiased, and so non-stationary data is differenced d times until the estimated autocorrelation function of $(1-T)^d y_t$ dies away fairly rapidly.

Having chosen d, it remains to choose p,q and estimate the parameters. Most methods assume normally-distributed errors (or are equivalent to doing so), and many are based on maximising the likelihood or an approximation to it. In general closed forms for the maximisation do not exist, so that numerical methods are used which usually give iterative solutions to the equations.

Alternative models can be compared by looking at maximum log - likelihoods. The principle of parsimony of Box-Jenkins is to only fit models of low order (p,q small). Procedures such as Akaike's AIC attempt to avoid over fitting models by subtracting from the maximised likelihood a constant multiplied by the dimension of the parameter vector.

Tests of model adequacy of stationary series examine whether p and q are appropriate. For example the Box-Pierce test looks at the first few autocorrelations of the residuals of the fitted model which is asymptotically chi-squared distributed.

Since we are concerned with forecasting it is worth quoting from Cox (1981, p 99)

"When there is a tightly specified objective, such as forecasting, considerations of choosing a notional 'true' model become less important and error of forecasting is the appropriate criterion for judging any particular model selection procedure."

We close this section with an example which not only gives an idea of the complexity involved in estimating parameters of the simplest models, but also shows a novel way of deriving this estimator.

Example 2.4

Consider the MA(1) model

$$y_t = \varepsilon_t + \beta \varepsilon_{t-1} \tag{2.19}$$

with β unknown. For a loss function $L(a,b)$, which we take to be a non-negative function $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with $L(a,a) \equiv 0$ (defined in Chapter 4) let us choose the value of β that minimises

$$\sum_{t=1}^{n} L(y_t, y_{t-1}(1)) \tag{2.20}$$

where n is the number of observations. That is we minimise the loss between the one-step ahead predictors and the observations.

From Example 2.3, if $y_0(1)=0$ is the initial predictor

then
$$y_{t-1}(1) = \beta(y_{t-1} - y_{t-2}(1))$$
$$= \beta y_{t-1} - \beta^2 y_{t-2} \ldots + (-1)^t \beta^{t-1} y_1 \quad (2.21)$$

so that on using the quadratic loss function $L(a,b)=(a-b)^2$ substituting (2.21) into (2.20), differentiating with respect to β and equating to zero gives

$$\sum_{t=2}^{n} (y_t - \beta y_{t-1} + \ldots + (-1)^{t-1}\beta^{t-1} y_1) \times$$
$$\times (y_{t-1} - 2\beta y_{t-2} \ldots + (-1)^{t-2}(t-1)\beta^{t-2} y_1) = 0.$$

That is
$$\sum y_t y_{t-1} + \sum_{r=1}^{n-1} \beta^r (-1)^r \{\sum y_{t-1} y_{t-r} + 2\sum y_{t-2} y_{t-r+1}$$
$$\ldots + (r+1) \sum y_{t-r-1} y_t\} = 0.$$

If we denote $r_k = \dfrac{\sum_{t=k}^{n} y_t y_{t-k}}{\sum y_t^2}$ then the condition simplifies to

$$r_1 - \beta(1 + r_2) + \beta^2(r_1 + 2r_1 + 3r_3) - \beta^3(r_2 + 2 + 3r_2 + 4r_4)$$
$$\ldots + (-1)^{\lambda-1}\beta^{n-1} \bar{x} = 0$$

or

$$r_1 - \beta(1 + r_2) + \beta^2(3r_1 + 3r_3) - \beta^3(2 + 4r_2 + 4r_4) + \ldots$$
$$+ (-1)^{n-1}\beta^{n-1} X = 0 \quad (2.22)$$

Where the last term $X = \{(n-2) + n(r_2 + r_4 + \ldots + r_{n-2})\}$, n-1 odd and $= n(r_1 + r_3 \ldots + r_{n-1})$ if n-1 even.

Assuming n-1 is odd, then on rearranging, the condition becomes

$$\beta\{1 + 2\beta^2 + 3\beta^4 + 4\beta^6 + \ldots + (n-2)\beta^{n-2}\}$$
$$= r_1\{1 + 3\beta^2 + 5\beta^4 + \ldots + (n-1)\beta^{n-2}\} +$$
$$+ r_3\{3\beta^2 + 5\beta^4 + \ldots\} - r_2\{2\beta + 4\beta^3 \ldots\}. \quad (2.23)$$

As n increases the term involving r_k tends to

$$(-1)^{k-1} \frac{\partial \beta^k}{\partial \beta(1-\beta^2)} = (-1)^{k-1} \frac{k(1-\beta^2)\beta^{k-1} + 2\beta^{k+1}}{(1-\beta^2)^2} .$$

Since $(1 + 2\beta^2 + 3\beta^4 \ldots) = \frac{1}{(1-\beta^2)^2}$

equation (2.23) simplifies to

$$\beta = \sum_{k=1}^{\infty} (-\beta)^{k-1} k r_k - 2\beta \sum_{k=1}^{\infty} (-\beta)^k r_k \qquad (2.24)$$

which is the same as the asymptotic likelihood equation in Godolphin (1977), but derived in a different way. This equation is well conditioned provided that the modulus of β is not too close to 1 so a starting value will converge to the single solution in the interval (-1,1). Note that formulae (2.22), (2.23) give exact recursions for finite samples using this approach. The method can be extended to looking at different loss functions, but in general the solution for β will be difficult.

CHAPTER 3

CONTROL THEORY

3.1 Introduction

In this chapter we give a brief discussion of the principles of control theory and introduce the fundamental concepts of controllability and observability. Much of the material for this chapter can be found in Kwakernaak and Sivan (1972), Kushner (1971), Jacobs (1974), Barnett(1975) and Gelb (1974). We shall only look at the discrete time case.

3.2 Control Systems - Non Stochastic

Consider a situation where a set of admissible inputs to a 'dynamical system' gives rise to observed outputs. We shall take the dynamic equations of the system to be

$$\underline{y}_i = g(\underline{x}_i, \underline{u}_i, i)$$

$$\underline{x}_{i+1} = f(\underline{x}_i, \underline{u}_i, i)$$

for functions f and g where the system behaviour is observed at discrete points t_i, i=0, 1, 2, ... This description also follows from Kalman's (1963b) axiomatic definition of a dynamical system. In general the $\underline{u}_i$ belong to the topological space of admissible inputs, $\underline{y}_i$ is the (m-dimensional) observation at time t_i and $\underline{x}_i$ is the state of the system, taking values in a topological space X. In what follows X will be real and finite dimensional Euclidean space - $\mathbb{R}^n$ say. Linear systems have the simpler description

$$\underline{y}_i = \underline{F}_i \underline{x}_i \tag{3.1}$$

$$\underline{x}_i = \underline{G}_i \underline{x}_{i-1} + \underline{B}_i \underline{u}_i \tag{3.2}$$

where $\underline{y}$, $\underline{x}$ and $\underline{u}$ are m, n and p vectors respectively with $\underline{F}_i$, $\underline{G}_i$, and $\underline{B}_i$ matrices of the appropriate dimension.

It is useful to note that systems of the form

$$\underline{y}_i = \underline{F}_i \underline{x}_i + \underline{C}_i \underline{u}_i$$

with the state evolving under (3.2) can be cast in the above form by defining an augmented state vector as

$$\underline{x}_i^* = \begin{pmatrix} \underline{x}_i \\ \underline{u}_i \end{pmatrix}$$

$\underline{F}_i^*$, $\underline{G}_i^*$, and $\underline{B}_i^*$ can then be obtained in a straightforward manner.

We shall often consider the following simplification

Definition 3.1

The system (3.1), (3.2) is said to be time invariant if the matrices $\underline{F}_i$, $\underline{G}_i$, $\underline{B}_i$ do not depend on i.

It is sometimes preferable to consider the form of the system in terms of the initial specification at some point $i=i_0$. This is achieved by solving the state difference equation (3.2) and substituting the result in (3.1). We present this formally as the following theorem and its corollary in the general and time invariant case.

Theorem 3.2

Equation (3.2) has the solution

$$\underline{x}_i = \underline{\Phi}(i,i_0)\underline{x}_{i_0} + \sum_{j=i_0+1}^{i} \underline{\Phi}(i,j)\underline{B}_j\underline{u}_j \qquad i \geq i_0+1$$

where $\underline{\Phi}(i,i_0)$, $i \geq i_0$ is the matrix

$$\underline{\Phi}(i,i_0) = \begin{cases} \underline{G}_i \cdots \underline{G}_{i_0+1} & i \geq i_0+1 \\ \underline{I} & i=i_0. \end{cases} \qquad (3.3)$$

The transition matrix $\underline{\Phi}$ is the solution to the homogeneous equation

$$\underline{\Phi}(i+1,i_0) = \underline{G}_{i+1}\underline{\Phi}(i,i_0) \qquad i \geq i_0$$

with $\underline{\Phi}(i_0,i_0) = \underline{I}$.

For the time invariant case $\underline{\Phi}(i,i_0) = \underline{G}^{i-i_0}$.

Corollary 3.3

The system (3.1), (3.2) has the solution

$$\underline{y}_i = \underline{F}_i\underline{\Phi}(i,i_0)\underline{x}_{i_0} + \underline{F}_i \sum_{j=i_0+1}^{i} \underline{\Phi}(i,j)\underline{B}_j\underline{u}_j \qquad (3.4)$$

The second term can be written as $\sum_{j=i_0}^{i} k(i,j)\underline{u}_j$

where $k(i,j) = \begin{cases} \underline{F}_{i_0}\Phi(i,j)\underline{B}_j & i>j \\ 0 & j=i_0 \end{cases} \qquad (3.5)$

is the pulse response matrix. For time invariant systems this is a function of i-j.

3.3 Stability

We now introduce some of the different forms of stability involved in deterministic control systems and comment on the connection between the different definitions.

The simplest form of stability seems to arise from the corresponding result for ordinary differential equations. The point 0 is an equilibrium point for the system $\underline{x}_i = \underline{f}(x_i, i)$ if $\underline{x}_k = 0$ for all $k \geq i_0$, for some i_0. Without loss of generality we can consider the point at the origin by defining new state variables to be deviations about a non-zero equilibrium point if necessary. We then have the following definitions for the equilibrium point x=0, putting $i_0 \equiv 0$.

Definition 3.4

x=0 is stable in the sense of Lyapunov if for every positive ε there exists a ∂ such that

$$\|x_0\| < \partial \text{ implies } \|x_k\| < \varepsilon \text{ for all } t_k \geq t_0.$$

Definition 3.5

x=0 is asymptotically stable if it is stable and $x_t \to 0$ as $t \to \infty$.

By only looking at linear systems

$$\underline{x}_{i+1} = \underline{G}_{i+1}\underline{x}_i \tag{3.6}$$

we can show that these definitions, which describe the behaviour we would like to see, are equivalent to the more manageable ones of Jazwinski (1970, chapter 7) involving the transition matrix Φ of (3.3). Defining the norm of an r x c matrix M = (mij) to be

$$\|M\|^2 = \sum_{i=i}^{r} \sum_{j=i}^{c} |mij|^2$$

we have

Theorem 3.6

For the system (3.6), Definitions 3.4 and 3.5 are equivalent to

(i) The system is stable if $\|\underline{\Phi}(k,0)\|$ is bounded above for all $t_k > t_0$

(ii) The system is asymptotically stable if in addition $\|\underline{\Phi}(k,0)\| \to 0$ as $t_k \to \infty$.

Proof

For (i) $\underline{x}_k = \underline{\Phi}(k,0)\underline{x}_0$, so $\|\underline{x}_k\| \leq \|\underline{\Phi}(k,0)\| \, \|\underline{x}_0\|$, so that (i) implies Definition 3.4 Conversely if (i) is not satisfied then we can choose an ε such that for all positive δ

$$\|x_0\| < \delta \text{ but } \|x_k\| \geq \varepsilon \text{ for all } t_k \geq t_0.$$

The equivalence of (ii) and 3.5 follows from the fact that $\underline{\Phi}(k,0) \to 0$ if and only if $\|\underline{\Phi}(k,0)\| \to 0$ if and only if $\|\Phi(k,0)x_0\| \to 0$.

A stronger type of asymptotic stability is given by

Definition 3.7

The system is uniformly asymptotically stable if $\|\underline{\Phi}(k,0)\| \leq c_1 \exp\{-c_2(t_k - t_0)\}$ for all $t_k \geq t_0$ and some fixed positive constants c_1 and c_2.

In the time invariant case where $\underline{G}_i = \underline{G}$ then

$$\underline{\Phi}(k,0) = \underline{G}^k.$$

But $\underline{G}^k \to 0$ as $k \to \infty$ if and only if the eigenvalues of $\underline{G}$ are less than one in modulus, proving

Theorem 3.8

In the time invariant system the origin is asymptotically stable if and only if the matrix $\underline{G}$ has eigenvalues less than one in modulus.

A slightly stronger result is given in Kwakernaak and Sivan (1972, page 454) namely

Theorem 3.9

The system $\underline{x}_i = \underline{G}\underline{x}_{i-1}$ is stable if and only if the eigenvalues of $\underline{G}$ have modulus less than one or equal to one with any modulus one eigenvalue of multiplicity m having m linearly independent eigenvectors.

For example the matrix $\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$ is stable whereas

$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ or $\begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix}$ are unstable.

Procedures for checking on whether the eigenvalues are less than one in modulus are the Routh-Hurwitz or Schur-Cohn criteria for example, which are explained in Jacobs (1974) or Jury (1964).

The above definitions relate to systems that do not have any inputs; we would like to know how our systems behave when inputs are applied. In particular control engineers like systems to behave 'nicely' - for example tend to a steady value - if an input is applied and then removed. They therefore use

Definition 3.10

A dynamical system is bounded-input bounded-output stable (b.i.b.o.) if its response to any bounded input is to produce a bounded output.

In fact we can show that the types of stability defined above do produce the required behaviour. For the linear case

Theorem 3.11

If $\underline{x}_i = \underline{G}_i\underline{x}_{i-1}$ is uniformly asymptotically stable then (3.1), (3.2) is b.i.b.o. stable if in addition $\|\underline{F}_i\|$ and $\|\underline{B}_i\|$ are bounded, say $\|\underline{F}_i\| < f$, $\|\underline{B}_i\| < b$, with f and b positive constants.

Proof

Taking norms in (3.4) with $i_0 \equiv 0$

$$\|\underline{y}_i\| \leq \|\underline{F}_i\| \{ \|\underline{\Phi}(i,0)\| \|\underline{x}_0\| + \sum_{j=1}^{i} \|\underline{\Phi}(i,j)\| \|\underline{B}_j\| \|\underline{u}_j\| \} \tag{3.7}$$

But from Definition 3.7, for some c_1, c_2

$\|\Phi(k,0)\| < c_1 \exp\{-c_2 k\}$ so if $\underline{u}_j$ is bounded, $\|\underline{u}_j\| < u$ for some u and

$$\|\underline{y}_i\| < f c_1 \{ e^{-c_2 i} \|\underline{x}_0\| + bu \sum_{j=1}^{i} e^{-c_2(i-j)} \}$$

which is bounded above as $\sum_{j=1}^{i} e^{-c_2(i-j)} < \frac{e^{-c_2}}{1-e^{-c_2}}$, thus

proving the theorem.

In the time invariant case we can prove the slightly stronger result.

Theorem 3.12

If $\underline{x}_i = \underline{G}\underline{x}_{i-1}$ is asymptotically stable, then

$$\underline{y}_i = \underline{F}\underline{x}_i$$

$$\underline{x}_i = \underline{G}\underline{x}_{i-1} + \underline{B}\underline{u}_i$$

is b.i.b.o. stable.

Proof

From (3.3), $\underline{\Phi}(i,0) = \underline{G}^i$, so in (3.7), if $\|\underline{u}_j\| < u$ for all j

$$\|\underline{y}_i\| < f\{\underline{G}^i\|\underline{x}_0\| + bu \sum_{j=1}^{i} \underline{G}^{j-i}\}.$$

But this is bounded above since the sequence $I+\underline{G} + \ldots + \underline{G}_n$ is convergent if $\underline{G}$ has eigenvalues less than one in modulus.

3.4 The z-transform

The z-transform is a useful tool in the analysis of discrete time systems, and is the discrete equivalent of the Laplace transform. For a discrete time vector variable $\underline{x}_i$, i = 0, 1 ... we have

Definition 3.13

The z-transform of $\{\underline{x}_i\}$ is given by

$$\underline{X}(z) = \sum_{i=0}^{\infty} \underline{x}_i z^{-i}$$

defined for those complex z for which $\underline{X}(z)$ is convergent. For a time invariant linear system, (3.1), (3.2) become

$$\underline{y}_i = \underline{F}\underline{x}_i \tag{3.8}$$

$$\underline{x}_{i+1} = \underline{G}\underline{x}_i + \underline{B}\underline{u}_{i+1}. \tag{3.9}$$

Because the Laplace transform of $\underline{x}_{i+1}$ is $zX(z) - z\underline{x}_0$, taking transforms of (3.9) gives

$$z\underline{X}(z) - z\underline{x}_0 = \underline{G}\underline{X}(z) + z\underline{B}\underline{U}(z) - z\underline{B}\,\underline{u}_0$$

where $\underline{U}(z)$ is the Laplace transform of u_i, and so rearranging

$$\underline{X}(z) = (z\underline{I}-\underline{G})^{-1}z\underline{B}\underline{U}(z) + (z\underline{I}-\underline{G})^{-1}z(\underline{x}_0-\underline{B}\underline{u}_0).$$

Taking transforms in (3.8) and substituting gives

$$\underline{Y}(z) = \underline{F}(z\underline{I} - \underline{G})^{-1}z\underline{B}\underline{U}(z) + \underline{F}(z\underline{I} - \underline{G})^{-1}z(\underline{x}_0 - \underline{B}\underline{u}_0)$$

which relates the input transform $\underline{U}$ to the output transform $\underline{Y}$.

The quantity $H(z) = \underline{F}(z\underline{I} - \underline{G})^{-1}z\underline{B}$ (3.10)

is called the z transfer matrix of the system.

For time invariant systems, the pulse response matrix (3.5) is $k_{ij} = F\underline{G}^{i-j}\underline{B}$; putting i-j = t, so that $k_t = \underline{F}\underline{G}^t\underline{B}$ then

$$H(z) = K(z),$$

that is H is the z-transform of the pulse response matrix.

The z-transform can be used to test stability as the following result, proved in Lindorff (1965) or Jury (1964) shows

Theorem 3.14

A linear discrete filter is b.i.b.o. stable if and only if its transfer function H(z) contains no poles on or outside the unit circle.

The proof of this theorem can also be seen from Theorems 3.8, 3.12 and equation (3.10). We shall illustrate the use of the above theorem by considering the Kalman filtering equation (which is explained in detail in the next chapter).

Example 3.15

In the steady state, the time invariant Kalman Filter has the form

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

where $\underline{A}$, $\underline{F}$, $\underline{G}$ are fixed matrices, $\underline{\theta}$ the 'state' of the system and $\underline{y}$ the observation vector. Taking transforms, using a notation consistent with definition (3.10),

$$\Theta(z) = (\underline{I}-\underline{AF})\underline{G}\ \frac{\Theta(z)}{z} + \underline{AY}(z)$$

so the z-transfer function is

$$H(z) = \{z\underline{I} - (\underline{I} - \underline{AF})\underline{G}\}^{-1} z\underline{A}.$$

The poles of $H(z)$ are the zeros of the determinant $|zI - (I - AF)G|$, which are the eigenvalues of $(I - AF)G$, and the system is stable if the poles of $H(z)$ lie inside the unit circle or the eigenvalues of $(\underline{I} - \underline{A}\,\underline{F})\underline{G}$ are less than one in modulus.

3.5 Observability and Controllability

Observability and controllability are key concepts in the theory of control theory. Losely speaking, controllability ensures that we can apply an input function which enable us to reach any given state, whilst observability means that the state vector can be calculated if we know the observations and the model equations. Used somewhat differently in a forecasting context, these two conditions will ensure that the filtering procedure we use has certain optimum properties and also that we use the model of the smallest dimension whilst retaining all of the information in the model. These concepts are developed in Chapters 6 and 7, with a foretaste in the following section.

Definition 3.16

The linear discrete time system (3.1), (3.2) is completely observable if for any t_0 and initial state $\underline{x}_0$ there exists a finite time $t_r > t_0$ such that knowledge of $\underline{u}_t$ and $\underline{y}_t$ for $t_0 \leq t \leq t_r$ is sufficient to determine $\underline{x}_0$.

Conditions to determine observability are given by

Theorem 3.17

The system (3.1), (3.2) is completely observable if and only if there exists a t_r such that

$$\underline{M}(r,0) = \sum_{i=0}^{r} \underline{\Phi}(i,0)^T \underline{F}_i^T \underline{F}_i \underline{\Phi}(i,0) > 0$$

that is the matrix $\underline{M}(r,0)$ is positive definite.

Proof

Consider the case of scalar y_t; without loss of generality assume that $u_t \equiv 0$, then

$$\begin{bmatrix} y_0 \\ . \\ . \\ . \\ y_r \end{bmatrix} = \begin{bmatrix} \underline{F}_0 \\ \underline{F}_1\underline{\Phi}(1,0) \\ . \\ . \\ . \\ \underline{F}_r\underline{\Phi}(r,0) \end{bmatrix} \underline{x}_0 . \qquad (3.11)$$

To obtain a solution for x_0 in terms of $(y_0 \ldots y_r)^T$ it is necessary and sufficient that the matrix premultiplying $\underline{x}_0$ has full rank. This is equivalent to requiring $\underline{M}(r,0)$ positive definite. The proof for vector y_t is similar. In fact since the matrix premultiplying $\underline{x}_0$ is an r+1 x n matrix, in the univariate case then $r \geq n-1$, and we can put $r = n-1$ in the statement of the theorem, ie $t_r = t_{n-1}$.

In the time invariant case we reserve the symbol M to describe the matrix appearing in (3.11), that is we define

$$\underline{M} = \begin{pmatrix} \underline{F} \\ \underline{F}\underline{G} \\ \underline{F}\underline{G}^{n-1} \end{pmatrix} \qquad (3.12)$$

corresponding to time points $t_0 \ldots t_{n-1}$. For this case the criterion for observability can be specified as follows

Theorem 3.18

The system (3.1), (3.2) is completely observable if and only if the observability matrix $\underline{M}$ has rank n.

As its name implies, controllability means that it is possible by the application of control functions to move the system to any state.

Definition 3.19

The system (3.1), (3.2) is completely controllable if for initial state $\underline{x}_0 = \underline{0}$ and final state $\underline{x}_f$ there is a finite time t_r and a control sequence $\underline{u}_t$, $t_0 \leq t \leq t_r$ such that $\underline{x}_r = \underline{x}_f$.

Again we can show that conditions to test for controllability are given by

Theorem 3.20

A necessary and sufficient condition for controllability is that there exists a t_r such that the symmetric matrix

$$\underline{W}(r,0) = \sum_{t=0}^{r-1} \underline{\Phi}(r,i+1)\, \underline{B}_{i+1} B_{i+1}^T\, \underline{\Phi}(r,i+1)^T$$

in positive definite. See for example Kwakernaak and Sivan (1972, page 460).

Indeed, $u_j = \underline{B}_j^T\, \underline{\Phi}(r,j)\underline{W}^{-1}\underline{x}_f$ is the requisite control.

In the time invariant case we have

Theorem 3.21

A necessary and sufficient condition for complete

controllability in the time invariant case is that the matrix

$$W = [B, GB \ldots G^{n-1}B] \qquad (3.13)$$

has rank n.

The two concepts of controllability and observability are in fact closely related through duality, namely

Theorem 3.22 Kwakernaak and Sivan (1972, page 466)

The system (3.1), (3.2) is completely controllable if and only if its dual system

$$\underline{y}^*_i = \underline{B}_{i^*-i}^T \underline{x}^*_i$$

$$\underline{x}^*_i = \underline{G}_{i^*-i}^T \underline{x}^*_{i-1} + \underline{F}_{i^*-i}^T \underline{u}^*_i$$

is completely observable, where i^* is an arbitrary fixed integer, and conversely.

Consequently any theorem for observability implies a corresponding controllability result, and vice-versa, for example Theorem 3.21 follows immediately from Theorem 3.18.

3.6 Algebraic Equivalence and Canonical Structure

In a vector space, we are free to choose our basis vectors, which gives rise to equivalent transformations or matrices, that is ones which have identical properties. Kalman (1963b) introduced the notation of algebraic equivalence for control systems.

Definitions 3.23

Two linear dynamic systems of the form (3.1), (3.2) with state vectors $\underline{x}$, $\underline{x}^*$ are algebraically equivalent whenever their phase vectors, defined by the pairs $(t,\underline{x})$, $(t,\underline{x}^*)$ are related for all t by $(t,\underline{x}^*) = (t,\underline{T}_t\underline{x})$ for some non singular matrix $\underline{T}_t$.

In other words there is a one-one correspondence between the phase spaces $T \times \underline{X}$ and $T \times \underline{X}^*$. In such a case the matrices $\underline{F}_t^*$, $\underline{G}_t^*$ and $\underline{B}_t^*$ are related to $\underline{F}_t$, $\underline{G}_t$ and $\underline{B}_t$ according to

$$\underline{F}_t^* = \underline{F}_t \underline{T}_t^{-1}$$

$$\underline{G}_t^* = \underline{T}_t \underline{G}_t \underline{T}_{t-1}^{-1}$$

$$\underline{B}_t^* = \underline{T}_t \underline{B}_t .$$

Algebraic equivalence is insufficient as it stands to preserve the stability properties of a linear dynamic system, for which we require topological equivalence. Topological equivalence is algebraic equivalence plus the additional conditions $\|\underline{T}_t\| \leq c_1$, $\|\underline{T}_t^{-1}\| \leq c_2$ for fixed constants c_1 and c_2. $\underline{T}_t$ can be varying even with time-invariant systems, however if it is constant then we use

Definition 3.24

Two constant linear dynamic systems are strictly equivalent if they are algebraically equivalent with $\underline{T}_t$ a constant matrix.

Using the Definition 3.23 with Theorems 3.17, and 3.20 we can show that

Theorem 3.25

Controllability and observability are preserved under algebraic equivalence.

It is then possible to prove the following canonical decomposition theorem.

Theorem 3.26

In a fixed linear dynamic system (3.1), (3.2) at every fixed instant of time there exists a co-ordinate system such

that the state variable can be decomposed into four mutually exclusive parts, $\underline{x}^T = (\underline{x}_{cu}, \underline{x}_{co}, \underline{x}_{uu}, \underline{x}_{uo})^T$ corresponding to parts which are completely controllable but unobservable, completely controllable and observable, uncontrollable and unobservable, and uncontrollable but completely observable respectively. Such decompositions always have the same number of state variables in each part, and it is possible to chose one such decomposition which produces the following canonical form

$$\underline{y}_t = \begin{pmatrix} 0 & F_t^B & 0 & F_t^D \end{pmatrix} \underline{x}_t$$

$$\underline{x}_t = \begin{pmatrix} G_t^{AA} & G_t^{AB} & G_t^{AC} & G_t^{AD} \\ 0 & G_t^{BB} & 0 & G_t^{BD} \\ 0 & 0 & G_t^{BC} & G_t^{CD} \\ 0 & 0 & 0 & G_t^{DD} \end{pmatrix} \underline{x}_{t-1} + \underline{B}_t \underline{u}_t$$

with

$$\underline{B}_t = \begin{pmatrix} B^A{}_t \\ B^B{}_t \\ 0 \\ 0 \end{pmatrix}$$

Example 3.27

Consider a time invariant system with observability matrix $\underline{M}$ found from (3.2). Let the non-singular matrix T be

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

where $\underline{T}_1$ is a basis for the subspace spanned by the rows of $\underline{M}$, and $\underline{T}_2$ is chosen to make $\underline{T}$ a basis for the n-dimensional state space, then defining $x_t^* = \underline{T}\underline{x}_t$ the system can be represented as

$$\underline{y}_t = (\underline{F}^*, 0)x_t^*$$

$$\underline{x}_t^* = \begin{pmatrix} \underline{G}_{11}^* & 0 \\ \underline{G}_{21}^* & \underline{G}_{22}^* \end{pmatrix} \underline{x}_{t-1}^* + \underline{B}_t^* \underline{u}_t$$

with $\{\underline{F}^* \ \underline{G}_{11}^*\}$ completely observable.

If the system is initially at rest, then from (3.4), (3.5)

$$\underline{y}_i = \sum_{j=0}^{i} \underline{k}(i,j)\underline{u}_j .$$

With a physical system it is possible to empirically determine the impulse response matrix $\underline{k}$ by applying a unit impulse to each input in turn. The question then arises as to whether such a matrix $\underline{k}(i,j)$ is realisable by a system of form (3.1), (3.2). Classical control theory uses difference equations

$$\underline{y}_i + \ldots + a_n \underline{y}_{i-n} = bo\, \underline{u}_i + \ldots + b_{n-1}\underline{u}_{i-n+1}. \qquad (3.13)$$

Taking Laplace transforms and ignoring the transient terms yields

$$\underline{Y}(z)\left[1 + \frac{\underline{a}}{z} + \ldots + \frac{\underline{a}_n}{z^n}\right] = \underline{u}(z)\left[bo + \frac{\underline{b}_1}{z} + \ldots + \frac{b_{\underline{n-1}}}{z^{n-1}}\right]$$

so in the notation of Section 3.3

$$\underline{H}(z) = \frac{b_0 z^n + \ldots + b_{n-1} z}{z^n + a_1 z^{n-1} + \ldots + a_n} \qquad (3.14)$$

Since for time invariant systems knowledge of $\underline{k}(i,j)$ is equivalent to knowledge of its z-transform $\underline{H}(z)$, then we say that realisations of a dynamical system expressed in terms of a difference equation (3.13), or by a time-invariant linear model (3.1), (3.2) are equivalent if they give the same z-transfer matrix.

To answer our question, it follows from (3.3) - (3.5) that we must have $\underline{k}(i,j) = \underline{P}(i)\underline{Q}(j)$ for some matrices P and Q. We say that a realisation is reducible of there is a proper subset of a realisation which also realises k. The main result is then

Theorem 3.28

Knowledge of the impulse response matrix identifies the completely controllable and completely observable part, and this part alone, of the dynamical system which generated it. This part is itself a dynamical system, has the smallest dimension among all realisations and is uniquely determined up to algebraic equivalence,
with the corollary,

Corollary 3.29

Every stationary impulse response matrix (that is a function of i-j) has constant irreducible realisations.

Example 3.30

In a time invariant univariate model m = p = 1, so from (3.10)

$$H(z) = \underline{F}(z\underline{I}-\underline{G})^{-1}z\underline{B}$$

But $(z\underline{I}-\underline{G})^{-1} = \text{adj.}(z\underline{I}-\underline{G})\ \det(z\underline{I}-\underline{G})^{-1}$ in which each term is a polynomical of degree n-1 divided by a polynomial of degree n. Consequently H(z) tends to a constant as z increases and so is of form

$$\frac{b_0 z^n + b_1 z^{n-1} + \ldots + b_{n-1} z}{z^n + a_1 z^{n-1} + \ldots + a_n}$$

the same form as (3.14). A realisation of such a model is given by

Theorem 3.31

A canonical decomposition of (3.14) is

$$\underline{F} = (b_{n-1} \cdots b_0)$$

$$\underline{G} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ 0 & \cdot & \cdot & & & 1 \\ -a_n & \cdot & \cdot & & & -a_1 \end{pmatrix} \qquad \underline{B} = \begin{pmatrix} 0 \\ \\ \\ \\ 0 \\ 1 \end{pmatrix}$$

provided that (3.14) has no common root.

Proof

By considering G of the above from and writing down $(z\underline{I}-\underline{G})^T$ or by an inductive argument we can show that

$$\text{adj.}(z\underline{I}-\underline{G})\underline{B} = \begin{pmatrix} 1 \\ z \\ \cdot \\ \cdot \\ z^{n-1} \end{pmatrix}$$

so that $H(z) = \underline{F}(z\underline{I}-\underline{G})^{-1}z\underline{B}$

$$= \underline{F}\ \text{adj.}(z\underline{I}-\underline{G})\ |\ z\underline{I}-\underline{G}\ |^{-1}z\underline{B}$$

where $|\ z\underline{I}-\underline{G}\ | = \det(z\underline{I}-\underline{G})$

$$= \frac{b_0 z^n + b_1 z^{n-1} + \ldots + b_{n-1}z}{|\ z\underline{I}-\underline{G}\ |}$$

But $|z\underline{I}-\underline{G}| = z^n + a_1 z^{n-1} + \ldots + a_n$ because $\underline{G}$ is a companion matrix, so that H(z) is as in (3.14). The decomposition is canonical if it is completely observable and completely controllable which is easily checked by applying the criteria of Theorems 3.18 and 3.21.

An alternative canonical decomposition, $\underline{F}^*, \underline{G}^*, \underline{B}^*$, is

$$\underline{F}^* = \underline{B}^T$$
$$\underline{G}^* = \underline{G}^T$$
$$\underline{B}^* = \underline{F}^T$$

with $\underline{F}$, $\underline{G}$, $\underline{B}$ as in Theorem 3.31.

The above process is analagous to converting an n^{th} order differential equation to a system of first order equations. For example for the alternative canonical decomposition above we are effectively defining a new state vector:

$$\underline{x}_i = \left(\begin{array}{c|cccc} 0 & -a_n & 0 & 0 & \cdots & 0 \\ 0 & -a_{n-1} & -a_n & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \\ 0 & -a_2 & -a_3 & \cdots & & -a_n \\ \hline 1 & 0 & \cdots & & & 0 \end{array}\right) \begin{pmatrix} y_i \\ y_{i-1} \\ \vdots \\ y_{i-n+1} \end{pmatrix} +$$

$$\left(\begin{array}{c|cccc} 0 & b_{n-1} & 0 & \cdots & 0 \\ & b_{n-2} & b_{n-1} & \cdots & 0 \\ & \vdots & & & \\ 0 & b & \cdots & & b_{n-1} \\ \hline 0 & 0 & & & 0 \end{array}\right) \begin{pmatrix} 0 \\ u_i \\ \vdots \\ u_{i-n+2} \end{pmatrix}$$

3.7 Stochastic Systems

We now let $\underline{u}_i$ be a discrete time vector stochastic process rather than a deterministic imput. If $\{\underline{u}_i\}$ is weakly or second order stationary, defined by analogy with §2.2, then let $\underline{\Gamma}_u(i-j)$ be the covariance matrix

$$\underline{\Gamma}_u(i-j) = E\{(\underline{u}_i - \underline{\mu})(\underline{u}_j - \underline{\mu})^T\}$$

where $\underline{\mu} = E(\underline{u}_i)$, and let $\underline{\Sigma}_u(\omega)$, $-\pi \leq \omega < \pi$ be the power spectral density matrix

$$\underline{\Sigma}_u(\omega) = \sum_{s=-\infty}^{\infty} e^{-i\omega s} \underline{\Gamma}_u(s). \qquad (3.15)$$

We then have the result, Kwakerwaak and Sivan (1972, page 469)

Theorem 3.32

If the input to an asymptotically stable time invariant linear discrete time system with z-transfer matrix $H(z)$ is weakly stationary then the output is weakly stationary. The output $\underline{y}$ has power spectral density matrix

$$\underline{\Sigma}_y(\omega) = H(e^{i\omega})\, \underline{\Sigma}_u(\omega) H^T(e^{-i\omega}) \quad -\pi \leq \omega \leq \pi. \qquad (3.16)$$

So for example if the input is a stationary Gaussian process, the output will be a stationary Gaussian process.

If we write $z \equiv e^{-i\omega}$ then (3.15) becomes with a slight abuse of notation

$$\underline{\Sigma}_u(z) = \sum_{-\infty}^{\infty} z^{-s} \Gamma_u(s)$$

the z-transform of the covariance matrix, and (3.16) becomes

$$\underline{\Sigma}_y(z) = \underline{H}(z) \underline{\Sigma}_u(z) \underline{H}(z^{-1}).$$

Example 3.33 considers the process.

$$y_{i+} \ldots + a_n y_{i-n} = b_o \varepsilon_i + \ldots + b_{n-1} \varepsilon_{n-i+1} \text{ as in (3.13)}$$

Then the process is asymptotically or bibo stable if the roots of $z^n + a_1 z^{n-1} \ldots + a_n$ lie within the unit circle (see eg Theorem 3.14), and the transfer function is given by (3.14). If the ε_i are a stationary Gaussian process - a

sequence of mutually uncorrelated random variables with constant variance matrix $\underline{C}$ say, then

$$\underline{\Sigma}_{\varepsilon}(z) = \underline{C}$$

and the output has covariance matrix with z-transform

$$\underline{H}(z)\underline{C}H(z^{-1})$$

The case of univariate y corresponds to the models of Chapter 2, and the remarks made concerning realisations and canonical representations of the last section apply. Indeed it follows from the above discussion that finding a model of form (3.1), (3.2) with stochastic input which realises a given wide sense stationary process is equivalent to finding a realisation of the transfer matrix $\underline{H}(z)$ of the system, which was discussed in the previous chapter.

We introduce the following two definitions which will be used in the analysis of stochastic systems in Chapter 4.

Definition 3.34

The system (3.1), (3.2) is uniformly completely observable if there is an integer $k \geq 1$ and positive constants αo, α_1, βo and β_1 such that

(a) $\underline{M}(r,r-k) > o$ for all r

(b) $\alpha_0 \underline{I} \leq \underline{M}^{-1}(r,r-k) \leq \alpha_1 \underline{I}$ for all r

(c) $\beta_0 \underline{I} \leq \underline{\Phi}(r,r-k)\,\underline{M}^{-1}(r,r-k)\,\underline{\Phi}^{T}(r,r-k) \leq \beta_1 \underline{I}$ for all r

where $\underline{M}$ and $\underline{\Phi}$ are defined in Theorems 3.17 and 3.2.

Definition 3.35

The dual condition is the system is uniformly completely controllable if there exists a $k \geq 1$ and positive constants $\alpha o, \alpha_1, \beta_o, \beta_1$ such that

(a) $\underline{W}(r+k,r) \geq 0$ for all r

(b) $\alpha_0 \underline{I} \leq \underline{W}^{-1}(r+k,r) \leq \alpha_1 \underline{I}$ for all r

(c) $\beta_0 \underline{I} \leq \underline{\Phi}^T(r+k,r)\underline{W}^{-1}(r+k,r)\,\underline{\Phi}(r+k,r) \leq \beta_1 \underline{I}$ for all r

where W is defined in Theorem 3.17.

In the time invariant case, systems are uniformly completely controllable/observable if and only if they are completely controllable/observable, so that the definitions in § 3.5 may be used.

CHAPTER 4

BAYESIAN FORECASTING

4.1 Introduction

The last chapter listed some of the key concepts in control theory, whose classical development was concerned with non-stochastic inputs and outputs. However, in the real world measurements are subject to error and for simplicity, or because of ignorance of all the relevant influences, our mathematical model will usually only be a good rather than a perfect description of reality. These problems lead naturally to the introduction of error terms, which can often be described probabilistically.

A natural question that arises is how to estimate the state in the presence of errors, or how to 'filter' out these quantities. For example control theory, including filtering theory, became very important during the American space programme, since one wanted to be able to estimate the position and velocity of space-vehicles - which correspond to state variables - in the presence of extraneous forces and measurement errors.

Harrison and Stevens in a series of papers (1971, 1975, 1976) used a similar methodology to attack problems in forecasting Time Series, which they termed Bayesian Forecasting. Such an approach differs from the classical approach and forms the starting point for the rest of this thesis.

Before embarking upon a brief description of Bayesian

forecasting we give a brief description of the ideas behind Bayesian inference.

4.2 Bayesian Inference

The foundations of Bayesian inference can be traced to the Reverend Thomas Bayes' work of 1763; a readable review of the subject is given in Lindley (1971) whose notation we now use. Other relevant works are Box and Tiao (1973), De Groot (1970), Ferguson (1967) and Raiffa and Schlaiffer (1961).

Suppose that we have a measure space of observations (X, Φ, ν) where ν is a σ-finite measure on the σ-field Φ of subsets of the general space X. Let $\{F_\theta, \theta\varepsilon\Theta\}$ be a family of probability measures on (X, Φ), each F_θ being dominated by ν so that

$$P(A|\theta) = \int_A p(x|\theta)\, d\nu(x) \qquad (4.1)$$

according to the Radon-Nikodym theorem, where A is any member of the σ-field Φ. $p(x|\theta)$ is thus a probability density function.

The Bayesian theory now assumes the existence of a probability system (Θ, Ω, P) where P is a probability measure on the σ-field Ω of subsets of Θ. If P is dominated by a measure μ on (Θ, Ω) then a prior density $p(\theta)$ can be defined on Ω. If further the likelihood function $p(x|\cdot)$ ($p(x|\theta)$ as a function of θ with x fixed) is P integrable and

$$\int_\Theta p(x|\theta)\, dP(\theta) = 0 \quad \text{implies that} \quad p(x|\theta) \equiv 0$$

then according to Bayes rule

$$p(\theta|x) = \frac{p(x|\theta)\ p(\theta)}{\int_\Theta p(x|\theta)\ p(\theta)\, d\theta}$$

provided that $p(x|\theta)$ is non-zero. Otherwise $p(\theta|x)=0$.

Manski (1981) mentions that such an analysis effectively imposes two restrictions. First, we require that the σ-finite measure μ has domain Ω, which implies no practical restrictions. But, secondly, the requirement that the likelihood be integrable means that it must be Ω-measurable, which can be severely restrictive in a practical sense. Manski discusses possible solutions to this problem, however we shall assume that the likelihood is integrable.

Equation (4.1) is a statement of conditional probability and a mathematical statement; disagreement arises over the nature and meaning of the prior, and whether or not the measure P exists. Cox and Hinkley (1974, p375) give three possible interpretations of the prior:

(a) As a frequency distribution; this might occur if the parameter is generated by a random mechanism amenable to statistical analysis.

(b) As an objective representation of what is rational to believe about a parameter - usually in the face of 'ignorance'.

(c) As a subjective probability assessment.

It is the last two cases that fall under the distinctly Bayesian umbrella. For the second case, various suggestions as to what constitutes a prior expressing ignorance have been made by Jeffreys (1961), Box and Tiao (1973) and more recently by Bernardo (1979) amongst others. Warnings against the use of improper priors in such situations have been made by Stone (1970) amongst others.

The third case can be expressed in decision theoretic terms - indeed it is a moot point as to whether or not all statistical theory involves decision theory. If we have a

decision space D of decisions d, and a real non-negative loss function $L(d,\theta)$ on $D \times \Theta$ then a Bayes decision is one that minimises the expected loss

$$E(L) = \int_{\Theta} L(d,\theta)\, dP(\theta)$$

provided that L is integrable and that a minimum exists. Some argue that it is more natural to work with bounded utility functions, $U(d,\theta)$ in which case a Bayes decision maximises the expected utility. Much work has centred on the use of convex loss functions, for example $L(d,\theta)=(d-\theta)^2$, when the Bayes decision is the expectation $E(\theta)$ with respect to $P(\theta)$. However it is well known that unbounded loss functions, such as the above, can create problems. A class of bounded loss functions has been given by Lindley (1976), whilst Smith (1980) has given bounds for decisions using bounded loss functions.

De Groot (1970) develops an axiomatic approach, considering the subjective probability space $(\Theta x X, \Omega x \Phi, F_{\theta}.P)$ and shows that the axioms, if accepted, lead to a unique choice of prior probabilities and utilities - unique that is for each person. Manski (1981) points out that this is effectively using Bayes Theorem, which says that a measure on a product space can be decomposed into marginal and conditional measures, rather than Bayes rule.

Smith (1978) adopts a pragmatic approach by requiring that if two priors $p_1(\theta)$, $p_2(\theta)$ are 'close' then their associated posteriors and decisions are 'close'. This then allows some latitude in the specification of priors. If we use the weak (or star) topology to define closeness, which is effectively the strongest requirement, then we require

(i) the likelihood is bounded

(ii) the set of discontinuities has measure zero with respect to the prior

the latter being redundant if we use absolutely continuous priors.

The choice of loss function plays an important role in practical situations, because under different loss functions the same posterior can give rise to very different decisions. In a sense this is because we are choosing a single point to summarise an entire distribution; indeed by choosing an appropriate loss function we can reach almost any decision. For example

Theorem 4.1

Let f be a probability density function with continuous derivative defined on the real line with boundary points l_1 and l_2 satisfying

(i) $\lim_{x \to l_1} f(x) = \lim_{x \to l_2} f(x) = 0$

(ii) f(x) is strictly unimodal with f'(x)=0 at one and only one point in (l_1, l_2), m say.

Then for each point $x \in (l_1, l_2)$ there is a loss function L increasing in (y-d), positive, with L(0)=0 such that the Bayes decision with respect to f and L is x.

Proof

Consider the asymmetric loss function of gauge (a,b) defined by

$$L_{a,b}(y-d) = \begin{cases} 0 & -a < y-d < b \qquad a,b>0 \\ 1 & \text{otherwise} \end{cases}$$

Then E(L) = 1 + F(d-a) - F(d+b) under the above conditions is a continuous function of d whose minimum satisfies

$$f(d-a) = f(d+b). \qquad (4.2)$$

At any point $x \in (l_1, l_2)$ $f(x)$ is non-zero; by the intermediate value theorem applied to (l_1, m) and (m, l_2) there are two points x_1 and x_2 such that

$$f(x_1) = f(x_2) = \frac{f(x)}{2}.$$

Then x is the Bayes decision for the loss function gauge $(x-x_1, x_2-x)$, because (4.2) holds, and x_1, x_2 are the only two points distance $|x_2-x_1|$ apart satisfying $f(x_1)=f(x_2)$. If this last condition does not hold then for some ε

$$f(x_1+\varepsilon) = f(x_2+\varepsilon);$$

but $x_1+\varepsilon < m < x_2+\varepsilon$ and $x_1 < m < x_2$, so that from (ii) $f(x_1+d) > f(x_1)$, $f(x_2+d) < f(x_2)$ so that $f(x_1+\varepsilon) \neq f(x_2+\varepsilon)$. Thus x is indeed the Bayes decision and the theorem is proved.

How one actually choses a loss function in a practical context when one is not estimating a parameter is a subject largely glossed over by Bayesians. If one accepts De Groot's axiomatic approach then it is a question of extracting the persons personal utilities, by asking suitable preference type questions. A more pragmatic approach might be to try and use an approximate bounded loss function with a small number of adjustable parameters that can be altered to approximately represent the individual's preferences. This whole question is in need of further research, and one that we shall from now on sidestep by concentrating on symmetric loss functions which provide a degree of 'impartiality'.

4.3 Bayesian Forecasting: The Dynamic Linear Model

The Bayesian Forecasting approach to Time Series analysis assumes the following description, known as a Dynamic Linear Model or DLM: an observation equation

$$\underline{y}_t = \underline{F}_t\underline{\theta}_t + \underline{v}_t \tag{4.3}$$

and a system equation

$$\underline{\theta}_t = \underline{G}_t\underline{\theta}_{t-1} + \underline{w}_t. \tag{4.4}$$

In general $\underline{y}_t$ is an m vector of observations, $\underline{\theta}_t$ an n vector of process parameters with $\underline{F}_t$, $\underline{G}_t$ m x n and n x n matrices respectively, known at time t. $\underline{v}_t$ and $\underline{w}_t$ are random normal vectors of appropriate dimension with

$$\underline{v}_t \sim N(0,\underline{V}_t) \tag{4.5}$$

$$\underline{w}_t \sim N(0,\underline{W}_t) \tag{4.6}$$

independently.

Such a description postulates an (unobserved) underlying process characterised by $\underline{\theta}_t$, which evolves in a Markov fashion, together with an observation equation relating the observations $\underline{y}_t$ to $\underline{\theta}_t$. For example one of the simplest models is the 'steady model',

$$y_t = \theta_t + v_t$$

$$\theta_t = \theta_{t-1} + w_t$$

which can be interpreted by saying that the observations are noisy measurements of an underlying level or mean, which is described by a random walk. Some consequences of this model are investigated in Chapters 5 to 7.

The theory for the model (4.3) - (4.4) with assumptions (4.5) - (4.6) is enshrined in the Kalman filter after Kalman (1963a).

Theorem 4.2

For a DLM with $\{\underline{F}_t, \underline{G}_t, \underline{V}_t, \underline{W}_t\}$ known at time t, if the posterior distribution for $\underline{\theta}_{t-1}$ at time t-1 is Normal

$$(\underline{\theta}_{t-1}|\underline{D}_{t-1}) \sim N(\underline{m}_{t-1}, \underline{C}_{t-1}) \tag{4.7}$$

then the posterior at time t is also Normal

$$(\underline{\theta}_t|\underline{D}_t) \sim N(\underline{m}_t, \underline{C}_t) \tag{4.8}$$

where $\underline{D}_t$ represents the data up to time t, so that

$$\underline{D}_t = (\underline{D}_{t-1}, \underline{y}_t, \underline{F}_t, \underline{G}_t, \underline{V}_t, \underline{W}_t).$$

The equations relating $\underline{m}_t$, $\underline{C}_t$ to $\underline{m}_{t-1}$, $\underline{C}_{t-1}$ are

$$\underline{m}_t = \underline{G}_t\underline{m}_{t-1} + \underline{A}_t\underline{e}_t \tag{4.9}$$

$$\underline{C}_t = \underline{P}_t - \underline{A}_t\hat{\underline{Y}}_t\underline{A}_t^T \tag{4.10}$$

where

$$\begin{aligned} \underline{e}_t &= \underline{y}_t - \hat{\underline{y}}_t \\ \hat{\underline{y}}_t &= \underline{F}_t\underline{G}_t\underline{m}_{t-1} \\ \underline{P}_t &= \underline{G}_t\underline{C}_{t-1}\underline{G}_t^T + \underline{W}_t \\ \hat{\underline{Y}}_t &= \underline{F}_t\underline{P}_t\underline{F}_t^T + \underline{V}_t \\ \underline{A}_t &= \underline{P}_t\underline{F}_t^T(\hat{\underline{Y}}_t)^{-1}. \end{aligned} \tag{4.11}$$

Proof

This can be found in any text-book on stochastic control theory, for example Jazwinski (1970, chapter 7). A particularly simple proof is given in Harrison and Stevens (1971), proved using Bayes rule

$$p(\underline{\theta}_t|\underline{D}_t) \propto p(\underline{y}_t|\underline{\theta}_t, \underline{F}_t, \underline{V}_t)\ p(\underline{\theta}_t|\underline{D}_{t-1})$$

where

$$p(\underline{\theta}_t|\underline{D}_{t-1}) = \int p(\underline{\theta}_t|\underline{\theta}_{t-1})\ p(\underline{\theta}_{t-1}|\underline{D}_{t-1})\ d\underline{\theta}_{t-1}\ .$$

Since all the relevant distributions are Normal, the relations simplify to give the above recursions.

Remarks

1. The quantities in (4.11) are of interest in themselves, for instance $\underline{\hat{y}}_t$, $\underline{\hat{Y}}_t$ are the expectation and variance of $\underline{y}_t$ conditional upon $\underline{D}_{t-1}$, so that $\underline{e}_t$ is the one step ahead forecasting error which occurs in many forecasting systems. $\underline{A}_t$ is the Kalman gain vector.

2. The process needs to start off with a prior for $\underline{\theta}_0$ so that the forecater needs to build his prior opinions into the model -gained from experience, forecasting in similar circumstances, inside knowledge and so on. The simplest case occurs when we can approximate prior knowlege by $p(\underline{\theta}_0) \sim N(\underline{m}_0, \underline{C}_0)$. In fact we can approximate any prior by a Normal mixture (see for example Sorenson and Alspach 1971) however in most practical cases the effect of the prior decays with time so that a single Normal distribution suffices.

3. If $\underline{V}_t$ is positive then

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

and

$$\underline{A}_t = \underline{C}_t \underline{F}_t \underline{V}_t^{-1} \qquad (4.12)$$

which are alternatives for (4.11).

We can allow a more general form for the error terms, for example

$$\begin{pmatrix} \underline{v} \\ \underline{w} \end{pmatrix} \sim N\left(\begin{bmatrix} \underline{\mu}_v \\ \underline{\mu}_w \end{bmatrix}, \begin{bmatrix} \underline{V} & \underline{R} \\ \underline{R}^T & \underline{W} \end{bmatrix} \right) \qquad (4.13)$$

with the suffix t understood, in which case

$$\underline{m}_t = \underline{G}_t \underline{m}_{t-1} + \underline{\mu}_w + \underline{A}_t \underline{e}_t \qquad (4.14)$$

$$\underline{C}_t = \underline{P}_t - \underline{A}_t \underline{\hat{Y}}_t \underline{A}_t^T \qquad (4.15)$$

where

$$
\begin{aligned}
\underline{\hat{y}}_t &= \underline{F}_t\underline{G}_t\underline{m}_{t-1} + \underline{F}_t\underline{\mu}_w + \underline{\mu}_v \\
\underline{e}_t &= \underline{y}_t - \underline{\hat{y}}_t \\
\underline{P}_t &= \underline{G}_t\underline{C}_{t-1}\underline{G}_t^T + \underline{W} \\
\underline{\hat{Y}}_t &= \underline{F}_t\underline{P}_t\underline{F}_t + \underline{F}_t\underline{R}^T + \underline{R}\underline{F}_t^T + \underline{V} \\
\underline{A}_t &= (\underline{P}_t\underline{F}_t^T + \underline{R}^T)\ (\underline{\hat{Y}}_t)^{-1}.
\end{aligned}
\tag{4.16}
$$

The above recursions can be thought of as defining posterior distributions, or as defining estimates $\underline{m}_t$, $\underline{C}_t$ of the mean and variance. In the latter case $\underline{m}_t$ is the Bayes decision for $\underline{\theta}_t$ within the class of loss functions $L(\underline{d},\underline{\theta}_t) = L_1(\underline{d} - \underline{\theta}_t)$ where L_1 is symmetric about zero and a non-decreasing function of $\| d-\theta \|$. For example with $L_1(x) = 0$ if $x=0$ and 1 otherwise, the Bayes estimate is the mode which for Gaussian random variables is the mean $\underline{m}_t$.

Jazwinski (1970, chapter 7) details several alternative derivations of the filter. For example the above filter is optimal for $\underline{m}_t$ in the sense of

(1) recursive least squares with appropriate weighting matrices corresponding to the variance matrices $\underline{V}$ and $\underline{W}$.

(2) The linear minimum variance estimator, that is a linear estimator chosen to minimise

$$
\begin{aligned}
E\{(\underline{\theta}_t-\underline{m}_t)^T\ (\underline{\theta}_t-\underline{m}_t)\} &= \text{trace } E\{(\underline{\theta}_t-\underline{m}_t)(\underline{\theta}_t-\underline{m}_t)^T\} \\
&= \text{trace } \underline{C}_t.
\end{aligned}
$$

In particular, if $\underline{v}_t$ and $\underline{w}_t$ are not Normally distributed, the filter is still optimal in this linear minimum variance sense. Whilst if $\underline{v}_t$ and $\underline{w}_t$ are Normal mixtures then the optimal non-linear filter is given by the Class II models of §4.7; see also Chapters 8 and 9.

The Kalman filter also provides a particularly

convenient framework in which to alter the model: provided that the posterior at time t-1 is normally distributed as say (4.7), and that (4.3), (4.4) are appropriate as we pass from time t-1 to t, then the posterior is as given in Theorem 4.2 . Consequently we can incorporate subjective information into the model in two ways: firstly we can alter the values of $\{\underline{F}_t, \underline{G}_t, \underline{V}_t, \underline{W}_t\}$ at time t to take account of new information; in the simplest such case we might express increased uncertainty by larger variances $\underline{V}_t$ and $\underline{W}_t$. Secondly we might decide to directly amend the posterior; for instance if we thought that there was a change in the system parameters, then we might model the posterior by $p(\underline{\theta}_{t-1}|\underline{D}_{t-1}) \sim N(\underline{\mu}^*_{t-1}, \underline{C}^*_{t-1})$ with typically $\underline{C}^*_{t-1} > \underline{C}_{t-1}$ to express increased uncertainty. An example is given in Harrison and Stevens (1976).

Models other than the steady model are given in Harrison and Stevens (1976). These include seasonal models and also the Markov polynomial models considered by Godolphin and Harrison (1975). In their case studies paper Harrison and Stevens only look at models whose non-seasonal part is the steady model or the linear growth model

$$y_t = \theta_t + \varepsilon_t$$

$$\theta_t = \theta_{t-1} + \beta_t + w_{1t}$$

$$\beta_t = \beta_{t-1} + w_{2t}.$$

This is readily put in DLM form and takes its name from the fact that the underlying level undergoes an increase β_t, which can be thought of as a growth term that pursues a random walk. Also the forecast function for such a model is linear.

A useful property of DLMs is the superposition property which says that a linear combination of linear models is itself a linear model. Again this is illustrated in Harrison and Stevens. The practical implication of this is that we can model the seasonal part, trend and other factors by DLMs and then combine them linearly to achieve a DLM for the whole process.

4.4 Forecasting with DLMs

One of the primary objectives of the theory is to make inferences about future observations, or about linear transformations of the observation vector. In Bayesian inference this requires knowledge of the predictive distributions. If we denote

$$\underline{m}_{k,t} = E(\underline{\theta}_{t+k}|\underline{D}_t) \tag{4.17}$$

$$\underline{C}_{k,t} = Var(\underline{\theta}_{t+k}|\underline{D}_t) \tag{4.18}$$

with $\underline{m}_{0,t} = \underline{m}_t$, $\underline{C}_{0,t} = \underline{C}_t$, using (4.4) we have

$$\underline{\theta}_{t+k} = \underline{G}_{t+k}\underline{\theta}_{t+k-1} + \underline{w}_{t+k} ;$$

conditioning on time t and using the normality of $\underline{\theta}_t|\underline{D}_t$ and $\underline{w}_t$ for all t gives

$$\underline{\theta}_{t+k}|\underline{D}_t \sim N(\underline{m}_{k,t} , \underline{C}_{k,t})$$

with parameters generated recursively from

$$\underline{m}_{k,t} = \underline{G}_{t+k}\underline{m}_{k-1,t} \tag{4.19}$$

$$\underline{C}_{k,t} = \underline{G}_{t+k}\,\underline{C}_{k-1,t}\underline{G}_{t+k}^{T} + \underline{W}_{t+k} \tag{4.20}$$

which requires knowledge of $\underline{G}_t$, $\underline{W}_t$ up to time t+k. Similarly defining

$$\underline{y}_{k,t} = E(\underline{y}_{t+k}|\underline{D}_t) \tag{4.21}$$

$$\underline{Y}_{k,t} = \mathrm{Var}(\underline{y}_{t+k}|\underline{D}_t) \tag{4.22}$$

then using (4.3) and taking conditional means and variances

$$\underline{y}_{k,t} = \underline{F}_{t+k}\underline{m}_{k,t} \tag{4.23}$$

$$\underline{Y}_{k,t} = \underline{F}_{t+k}\underline{C}_{k,t}\underline{F}_{t+k}^{T} + \underline{V}_{t+k} \tag{4.24}$$

and the predictive distributions are Normal:

$$\underline{y}_{t+k}|\underline{D}_t \sim N(\underline{y}_{k,t}, \underline{Y}_{k,t}).$$

In particular $\underline{y}_{1,t} = \hat{\underline{y}}_t$, $\underline{Y}_{1,t} = \hat{\underline{Y}}_t$. Note that in the time invariant case

$$\underline{y}_{k,t} = \underline{F}\underline{G}^{k}\underline{m}_t \ . \tag{4.25}$$

These results have been derived under assumptions (4.5) and (4.6). Under assumptions (4.13) the formulae are slightly altered,

$$\underline{m}_{k,t} = \underline{G}_{t+k}\underline{m}_{k-1,t} + \underline{\mu}_{w,t+k} \tag{4.26}$$

$$\underline{C}_{k,t} = \underline{G}_{t+k}\underline{C}_{k-1,t}\underline{G}_{t+k}^{T} + \underline{W}_{t+k} \tag{4.27}$$

$$\underline{y}_{k,t} = \underline{F}_{t+k}\underline{m}_{k,t} + \underline{\mu}_{v,t+k} \tag{4.28}$$

$$\underline{Y}_{k,t} = \underline{F}_{t+k}\underline{C}_{k,t}\underline{F}_{t+k}^{T} + \underline{F}_{t+k}\underline{R}_{t+k}^{T} + \underline{R}_{t+k}\underline{F}_{t+k}^{T}. \tag{4.29}$$

Frequently (4.23) or (4.28) are used as k-step ahead predictors of $\underline{y}_{t+k}$; under the Normality assumptions the marginal predictive distributions are themselves Normal, so such forecasts are optimal under the wide class of symmetric loss functions mentioned in §4.3. As has already been mentioned, in a practical situation we might want to use different loss functions such as asymmetric ones to represent the differing consequences of overestimating or underestimating.

The following is another instance of when such forecasts might not be the best.

Example 4.3

Suppose that we have univariate data and make a transformation $z_t = \log y_t$. This is common practice in Time Series Analysis for various reasons, for example we might think that a multiplicative model is more appropriate than an additive one. Suppose further that z_t can be modelled as a univariate DLM, so that F, G in (4.3) and (4.4) are scalars. Then applying the Kalman filter will produce the predictive distribution

$$z_{t+k}|z^t \sim N(z_{k,t}, Z_{k,t}) . \qquad (4.30)$$

In classical analysis $z_{k,t}$ is taken as the forecast of z_{t+k} and so $\exp(z_{k,t})$ is used as the forecast of y_{t+k}. This corresponds to using a symmetric loss function on (4.30). However the primary quantities of interest are the observations y_{t+k}. Now the conditional distribution (4.30) corresponds to to log y_{t+k} having a (conditional) normal distribution, so that $y_{t+k}|y^t$ has a log-normal distribution with density function

$$\frac{1}{\sqrt{2\pi Z_{k,t}}} \frac{1}{y_{t+k}} \exp\{-\tfrac{1}{2} \frac{(\log y - z_{k,t})^2}{Z_{k,t}}\} .$$

If we apply the quadratic loss function to this quantity then the Bayes estimate, the mean

$$\exp\{ z_{k,t} + \tfrac{1}{2} Z_{k,t} \},$$

is the 'usual' estimate multiplied by $\exp(\tfrac{1}{2} Z_{k,t})$, which can take values considerably different from 1. It is more sensible to apply loss functions to the quantities of interest, and in this example different symmetric loss functions will produce different answers. In certain cases

the traditional estimate $\exp(z_{k,t})$ might not be very sensible.

An example that will recur throughout the thesis in various guises is the following steady model mentioned in § 4.3

Example 4.4

$$y_t = \theta_t + v_t \tag{4.31}$$

$$\theta_t = \theta_{t-1} + w_t \tag{4.32}$$

where all the quantities are univariate. For this model $F = G = 1$, and if $v_t \sim N(0, V)$, $w_t \sim N(0,W)$ with constant variance error terms then the Kalman filter equations (4.8) - (4.11) reduce to

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

$$A_t = \frac{C_{t-1} + W}{C_{t-1} + W + V} \qquad C_t = A_t V$$

and the predictive distributions (4.23), (4.24) give

$$y_{t+k} | y^t \sim N(m_{t+k}, C_t + kW + V)$$

so that under symmetric loss $y_t(k) = y_t(1) = m_t$ and the model produces constant forecasts.

Either by applying Corollary 4.6 to the model (4.31), (4.32) or directly we have that A_t tends to a limit A, so that in the limit $A = \{-W + (W^2+4VW)^{\frac{1}{2}}\}/2V$, and

$$y_t(k) = y_t(1) = m_t \tag{4.33}$$

where
$$m_t = m_{t-1} + A (y_t - m_{t-1}). \tag{4.34}$$

But (4.34) has the solution

$$m_t = A \sum_{j=0}^{\infty} (1-A)^j y_{t-j} \tag{4.35}$$

which is the same as (2.15) if $A=1+\beta$. Consequently this model produces the same forecasts as an ARIMA(0,1,1) model. In Chapters 6 and 7 we generalise this result to predictor equivalence between a wide class of DLMs and ARIMA models.

In Chapter 5 we show how even with steady models, if non-normal models and differing loss functions are used then (4.34) can be violated.

4.5 Applications of Control Theory

The error covariance term $\underline{C}_t$ is calculated recursively from (4.10), (4.11) as

$$\underline{C}_t = \underline{G}_t\underline{C}_{t-1}\underline{G}_t^T + \underline{W}_t - (\underline{G}_t\underline{C}_{t-1}\underline{G}_t^T + \underline{W}_t)\underline{F}_t^T \quad x$$
$$x\ (\underline{F}_t\underline{G}_t\underline{C}_{t-1}\underline{G}_t^T\underline{F}_t^T + \underline{F}_t\underline{W}_t\underline{F}_t^T + \underline{V}_t)^{-1}\ \underline{F}_t \quad x$$
$$x\ (\underline{G}_t\underline{C}_{t-1}\underline{G}_t^T + \underline{W}_t) \qquad (4.36)$$

This equation is independent of the data and so can be calculated as soon as the quantities $\{\underline{F}_t\ \underline{G}_t\ \underline{V}_t\ \underline{W}_t\}$ are known. In certain cases this matrix Ricatti equation tends to a limit $\underline{C}$. Equivalently, since

$$\underline{C}_t = (\underline{I} - \underline{A}_t\underline{F}_t^T)\underline{P}_t$$
$$\underline{P}_t = \underline{G}_t\underline{C}_{t-1}\underline{G}_t^T + \underline{W}_t$$

it can be seen that $\underline{P}_t$ tends to a limit if and only if $\underline{C}_t$ tends to a limit. The updates for $\underline{P}_t$ are sometimes simpler to work with; they are

$$\underline{P}_t = \underline{G}_t\underline{P}_{t-1}\underline{G}_t^T + \underline{W}_t - \underline{G}_t\underline{P}_{t-1}\underline{F}_{t-1}^T(\underline{F}_{t-1}\underline{P}_{t-1}\underline{F}_{t-1}^T + \underline{V}_{t-1})^{-1}\ x$$
$$x\ \underline{F}_{t-1}\underline{P}_{t-1}\underline{G}_t^T \qquad (4.37)$$

It turns out that sufficient conditions are those of observality and controllability given in Chapter 3. Observability implies that a solution exists and controllability that such a solution is unique. In complete generality we have from Kwakernaak and Sivan (1972, p535 theorem 6.45) and Jazwinski (1970, p240 theorem 7.4)

Theorem 4.5

If $\underline{F}_t$, $\underline{G}_t$, $\underline{V}_t$ and $\underline{W}_t = \underline{M}_t\underline{W}_t^*\underline{M}_t^T$ are bounded for all t and $\underline{W}_t^* \geq \alpha\underline{I}$, $\underline{V}_t \geq \beta\underline{I}$ for all t , for positive constants α, β then

(i) if the DLM is either observable or uniformly asymptotically stable with $\underline{C}_0=0$ ($\underline{P}_0=0$) then the variance $\underline{C}_t$ ($\underline{P}_t$) tends to a steady state solution of (4.36) {(4.37)} as $t\to\infty$.

Moreover if the system

$$\underline{\theta}_t = \underline{G}_t\underline{\theta}_{t-1} + \underline{M}_t\underline{W}_t^*$$

$$\underline{y}_t = \underline{F}_t\underline{\theta}_t$$

is either both uniformly completely observable and uniformly completely controllable or uniformly asymptotically stable then $\underline{C}_t$ ($\underline{P}_t$) tends to a unique solution of (4.36) (or (4.37)) for all initial conditions. Also the filter is uniformly asymptotically stable; that is the filter

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

The relevant definitions are all in Chapter 3. This theorem (part(ii)) is important because it means that not

only does a steady state exist which means that the prior effects decay to zero, but also it ensures that the recursions for $\underline{C}_t$ and $\underline{P}_t$ are numerically stable, which is important for computational purposes.

We shall mainly look at time-invariant sytems for which observability and controllability imply uniform observability and controllability, so that the theorem can be expressed more simply as

Corollary 4.6

If the time-invariant system

$$\underline{y}_t = \underline{F\theta}_t + \underline{v}_t \qquad \underline{v}_t \sim N(0,\underline{V})$$

$$\underline{\theta}_t = \underline{G\theta}_{t-1} + \underline{w}_t \qquad \underline{w}_t \sim N(0,\underline{W})$$

is observable then $\underline{C}_t$ ($\underline{P}_t$) converges to a limit provided that $\underline{C}_0=0$ ($\underline{P}_0=0$). If in addition $\underline{W}=\underline{MM}^T$ with ($\underline{G}$, $\underline{M}$) controllable then $\underline{C}_t$ ($\underline{P}_t$) converges to the unique solution of (4.36) ((4.37)) irrespective of the initial conditions.

Now from (3.13), ($\underline{G}$, $\underline{M}$) is controllable if

$$\underline{C} = (\underline{M}, \underline{GM}, \ldots, \underline{G}^{m-1}\underline{M})$$

has rank m where $\underline{G}$ is m x m and $\underline{M}$ is m x p. But if $\underline{W}$ is positive definite then there is an m x m matrix of full rank such that $\underline{W} = \underline{MM}^T$, so that $\underline{C}$ has rank m. Thus in the time invariant case, provided that $\underline{W}$ is positive definite, Corollary 4.6 will be satisfied if the system is observable.

In Chapter 6 it will be shown that it is sufficient to look only at observable systems, in which case provided that the system covariance matrix is positive definite

(which is usually true) the filter is stable and the covariance matrices $\underline{C}_t$, $\underline{P}_t$ will converge.

The algebraic equivalence of Definition 3.23 can be applied to the model (4.3), (4.4) relating the state vectors $\underline{\theta}_t$ and $\underline{\theta}_t^*$ by a non-singular matrix $\underline{T}_t$ via

$$\underline{\theta}_t = \underline{T}_t\underline{\theta}_t^* .$$

The matrices $\underline{F}_t^*$, $\underline{G}_t^*$ of the algebraically equivalent system are related to $\underline{F}_t$, $\underline{G}_t$ by

$$\underline{F}_t^* = \underline{F}_t\underline{T}_t^{-1}$$

$$\underline{G}_t^* = \underline{T}_t\underline{G}_t\underline{T}_{t-1}^{-1}$$

and the error terms $\underline{w}_t^* \sim N(0,\underline{W}_t^*)$ by

$$\underline{w}_t^* = \underline{T}_t\underline{w}_t$$

so

$$\underline{W}_t^* = \underline{T}_t\underline{W}_t\underline{T}_t^T .$$

The Kalman updates (4.9) -(4.11) of the two systems are related by

$$\underline{m}_t^* = \underline{T}_t\underline{m}_t \tag{4.38}$$

$$\underline{C}_t^* = \underline{T}_t^{-1}\underline{C}_t\underline{T}_t^T \tag{4.39}$$

with

$$\underline{A}_t^* = \underline{T}_t\underline{A}_t$$

$$\underline{P}_t^* = \underline{T}_t\underline{P}_t\underline{T}_t^T$$

and

$$\hat{y}_t^* = \hat{y}_t \tag{4.40}$$

$$\underline{\hat{Y}}_t^* = \underline{\hat{Y}}_t \tag{4.41}$$

provided that the priors are related by (4.38),(4.39). In fact we can show more generally than (4.40),(4.41) that the recursions for $\underline{y}_{k,t}$, $\underline{Y}_{k,t}$ of (4.21), (4.22) are

preserved under algebraic equivalence. This means that the predictive distributions (for the observations) are identical for any algebraically equivalent models, thus enabling us to use equivalent models.

4.6 Class I models

The Kalman Filter requires knowledge of $\underline{F}$, $\underline{G}$, $\underline{V}$ and $\underline{W}$ at time t. In practice it is extremely unlikely that we know these quantities exactly, especially in statistical applications where no physically based model presents itself. This problem is discussed further in Chapter 8 however one solution is to use the Harrison-Stevens Class I model. Such an approach assumes that during the time interval under consideration the data is best described by a single but unknown model M_i, where M_i is one of n possible models, i=1...n, each M_i corresponding to a set of values for $\{\underline{F}_t, \underline{G}_t, \underline{V}_t, \underline{W}_t\}$ throughout the time period.

If $p_0=(p_{1,0} \cdots p_{n,0})$ is the vector of prior probabilities whose j^{th} term is the probability that model j is 'correct' at time t=0, then defining

$$p_{j,t} = P(M=M_j | y^t, p_0) \qquad (4.42)$$

and using Bayes theorem

$$p_{j,t} \propto P(y_t | M_j, y^{t-1})\ p_{j,t-1} \cdot$$

The first term in the product is the likelihood. If we denote the probability density of a normal random variable mean μ variance σ^2 evaluated at x by $N_{\sigma^2}(x-\mu)$ then

$$P(y_t | M_j, y^{t-1}) = N_{\hat{Y}_j}(y_t - \hat{y}_j)$$

where $\hat{y}_j$, $\hat{Y}_j$ are the mean and variance of the predictive

distribution $p(y_t|y^{t-1})$ calculated from the Kalman filter assuming that M_j is in operation from time $t = 0$ to t. Consequently

$$p_{j,t} = \frac{p_{j,t-1}\, N_{\hat{Y}_j}(y_t - \hat{y}_j)}{\sum_{i=1}^{n} p_{i,t-1}\, N_{\hat{Y}_i}(y_t - \hat{y}_i)} \cdot \tag{4.43}$$

This is the prior-posterior analysis for the probability that model j is correct. Once we have these updates, the remaining quantities of interest are readily calculated, for example

$$\begin{aligned} p(\underline{\theta}_t|y^{t-1}) &= \sum p(\underline{\theta}_t|y^{t-1},M_i)\, p_{i,t-1} \\ p(\underline{\theta}_t|y^{t}) &= \sum p(\underline{\theta}_t|y^{t}, M_i)\, p_{i,t} \\ p(y_{t+k}|y^{t}) &= \sum p(y_{t+k}|y^{t},M_i)\, p_{i,t} \end{aligned} \tag{4.44}$$

where for instance $p(\underline{\theta}_t|y^t,M_i)$ is the posterior for $\underline{\theta}_t$ if M_i is in operation for all time and is found from the Kalman filter by using the values of $\{\underline{F},\underline{G},\underline{V},\underline{W}\}$ which correspond to M_i. In other words at each time stage we apply the Kalman filter n times, once for each set of values from model i, yielding n normal distributions which we then mix by using the appropriate probabilities. Decisions are then based upon these normal mixtures.

Such models might be used

(i) as a discrete form of Bayesian parameter estimation. In this case each M_i will correspond to a particular set of parameter values, and $p_{i,t}$ represents the posterior probability that the parameters have values

as in model i. For example if the variances V and $\underline{W}$ are unknown then we could set up a grid of values to cover their possible ranges. Some care needs to be exercised in the choice of this grid, and the implications of the simplications inherent in this method have not been studied.

(ii) If we are unsure as to which of a number of models best describes the data then we can use this method to choose a single model or a smaller number of models.

(iii) In situation (ii) we can use the forecasts from the whole suite of models, possibly to cater for parameter values changing in time, so that this is using a convex combination of models. If in fact the data are generated from a single model, such an approach will lead to a loss of performance.

(iv) If there is little or no data available, so that classical approaches to analysis are inappropriate. By using the class I procedure we can build upon our prior opinions.

4.7 Class II Models

In some applications of Time Series, expecially those involving economic data, it is extremely unlikely that we shall be able to successfully model any particular series by a model which has the same dynamics for all time. One possibility is to postulate a class of models at each time stage and allow the process to jump between models. The simplest way of describing the jumps is to specify a Markov evolution with a transition matrix $\{\pi_{ij}(t)\}$, π_{ij} being the probability that model j is operative at time t given

that model i operates at time t-1. This very general structure can be simplified by having $\{\pi_{ij}(t)\}$ independent of t, and simplified further by having $\pi_{ij} = \pi_j$, with $\pi_j \geq 0$, $\sum \pi_j = 1$, which makes $\pi_{ij}(t)$ independent of the past. These simplifications were assumed by Harrison-Stevens (1975, 1976), although relaxing these assumptions introduces a much wider class of models and it is envisaged that further research could lead to practical examples of such models. Note that $\pi_{ij} = \delta_{ij}$ gives the class I model of §4.6.

Suppose that at time t-1 the posterior distribution is a mixture of m normal distributions corresponding to m past histories $H_1 \ldots H_m$, for example each past history might correspond to any one of n models being involved at each time stage. If n models $M_1 \ldots M_n$ are introduced at time t then using the transition matrix $\{\pi_{ij}\}$ we can calculate the m x n transition matrix giving the transition probabilities from H_i to M_j, which without ambiguity we can denote as (π_{ij}). At time t - 1 the posterior distribution can be represented by

$$p(\underline{\theta}_{t-1}|y^{t-1}) = \sum_{i=1}^{m} p_{i,t-1} \; N_{\underline{C}_{i,t-1}}(\underline{\theta}_{t-1} - \underline{m}_{i,t-1}) \qquad (4.45)$$

where $p_{i,t-1}$ can be thought of as the probability that model i is operative at time t - 1.

Now

$$p(\underline{\theta}_t|y^{t-1}) = \sum_{i,j} p(\underline{\theta}_t|y^{t-1},M_j,H_i)\; p(M_j,H_i|y^{t-1}) \qquad (4.46)$$

and so $p(\underline{\theta}_t|y^{t-1}) = \sum_{i,j} p(\underline{\theta}_t|y^{t-1}\, M_j,H_i)\pi_{ij}\; p_{i,t-1}$

where $p(\underline{\theta}_t|y^{t-1},M_j,H_i)$ is calculated from the Kalman filter - in fact it is normal with mean $\underline{G}_j\underline{m}_{i,t-1}$ and covariance

matrix $\underline{G}_j \underline{C}_{i,t-1} \underline{G}_j^T + \underline{W}_j$. By similarly conditioning

$$p(\underline{\theta}_t | y^t) = \sum_{i,j} p(\underline{\theta}_t | y^t, M_j, H_i)\, p(M_j, H_i | y^t) \qquad (4.47)$$

where again $p(\underline{\theta}_t | y^t, M_j, H_i)$ is the normal distribution calculated from the updating procedure using the parameter values of model M_j and history H_i. We shall denote the mean and covariance matrix of this distribution by $\underline{m}_t(i,j)$ and $\underline{C}_t(i,j)$ respectively. Finally

$$p_t(i,j) \triangleq p(M_j, H_i | y^t) \propto p(y_t | M_j, H_i, y^{t-1})\, p(M_j, H_i | y^{t-1})$$

$$= \frac{N_{\hat{Y}_t(i,j)} \{y_t - \hat{y}_t(i,j)\}\, \pi_{ij}\, p_{i,t-1}}{\sum_{i,j} N_{\hat{Y}_t(i,j)} \{y_t - \hat{y}_t(i,j)\} \pi_{ij}\, p_{i,t-1}} \qquad (4.48)$$

substituting (4.48) into (4.47) gives the posterior density and we can similarly calculate the predictive distributions using the above probabilities.

The problem is that we now have a mn component normal mixture instead of an n component mixture, so that the situation is 'explosive'. Harrison-Stevens introduce a collapsing procedure to overcome this by reducing the mn components to n. This is achieved by integrating over the past histories for each model M_j, giving the following relations:

$$p_{j,t} = \sum_{i=1}^{m} p_t(i,j)$$

$$\underline{m}_{j,t} = \frac{\sum_{i=1}^{m} p_t(i,j)\, \underline{m}_t(i,j)}{p_{j,t}} \qquad (4.49)$$

$$\underline{C}_{j,t} = \frac{\sum_{i=1}^{m} p_t(i,j)\ \underline{C}_t(i,j) + \{\underline{m}_t(i,j) - \underline{m}_{j,t}\}\{\underline{m}(i,j) - \underline{m}_{j,t}\}^T}{p_{j,t}}$$

giving rise to new posterior for θ_t,

$$p(\underline{\theta}_t | y^t) = \sum_{i=1}^{n} p_{i,t-1} N_{\underline{C}_{i,t}}(\underline{\theta}_t - \underline{m}_{i,t}).$$

The above relations are obtained by equating the first two movements of the collapsed and uncollapsed systems.

The class II models admit another interpretation if $\pi_{ij} = \pi_j$ and the n different models consist of different values of the observation and state noise variances. Then we are effectively modelling non-normal error terms, namely mixtures of normal distributions,

$$v_t \sim \sum p_i N_{v_i}(v_t).$$

Conversely given error terms that are mixtures of normal distributions we can model them in this way. This is extremely useful since we can approximate any distribution by a normal mixture (see Sorenson and Alspach, 1971) and thus use any noise distribution.

4.8 Parameter Estimation

The Kalman updating procedure (4.12 - 4.16) assumes that the matrices $\underline{F}$, $\underline{G}$, $\underline{V}$ and $\underline{W}$ are known at time t. Indeed it is necessary to know or have estimates of these inputs to the filter if we are to be able to derive forecasts. In practice some or all of the matrices will be unknown and so estimates are required. As we have already mentioned, one solution is to use the Class I models defined above by putting a discrete grid of values on the unknown parameter, and then updating the posterior probabilities on each of these values. However we now consider different approaches to the problem.

Classical Time-Series analysis is very much concerned with the identification and fitting of models to particular series, where it is assumed that some single model (possibly non-linear and time-dependent) can adequately describe the data. We shall show later in Chapters 6 and 7, particularly Theorem 6.13 and following, that constant DLMs are equivalent in a suitably defined sense to ARIMA models. Consequently techniques applicable to ARIMA models can be used to identify and estimate the parameters of an ARIMA model, from which an equivalent state-space form can be used as a DLM. This is not a particularly interesting procedure because the state-space form is almost redundant when so used. We are more concerned with situations where interpretability of the model is important and also where it is inappropriate to postulate a single model valid for all time, in which case classical procedures are of less value. This point is developed in Chapters 7 and 8.

The underlying equivalence also implies that classical procedures can be applied directly to DLMs, for example

Maximum Likelihood Estimates

For univariate observations y_t, with $\underline{F}$, $\underline{G}$ known but V, W not, then

$$y_t = \underline{F}\,\underline{\theta}_t + v_t \qquad v_t \sim N(0,V) \tag{4.50}$$

$$\underline{\theta}_t = \underline{G}\,\underline{\theta}_t + \underline{w}_t \qquad \underline{w}_t \sim N(0,W). \tag{4.51}$$

By repeated application of (4.51)

$$\underline{\theta}_r = \underline{G}^r\underline{\theta}_0 + \underline{G}^{r-1}\underline{w}_1 + \ldots + \underline{w}_r \,. \tag{4.52}$$

If we denote the vector of n observations y_i by $\underline{y} = (y_1 \ldots y_n)^T$, then if $\underline{\theta}_0$ is known, the log-likelihood is given from (4.50), (4.51) as

$$L(V, \underline{W}) = -\frac{n}{2}\log(2\pi) - \tfrac{1}{2}\log|\underline{\Sigma}| - \tfrac{1}{2}(\underline{y}-\underline{\mu})^T\underline{\Sigma}^{-1}(\underline{y}-\underline{\mu}) \tag{4.53}$$

where $\underline{\mu}^T = \underline{F}(\underline{G} \;\; \underline{G}^2 \ldots \underline{G}^n)\,\underline{\theta}_0$ and

$$\underline{\Sigma} = (\Sigma_{ij}), \quad \Sigma_{ij} = \delta_{ij}V + \underline{F}\sum_{k=1}^{i}\underline{G}^{i-k}\underline{W}(\underline{G}^T)^{j-k}\underline{F}^T$$

and where δ is the Kronecker delta; $\underline{\mu}$, $\underline{\Sigma}$ are the mean and covariance matrices of the observations respectively.

If instead $\underline{\theta}_0$ is unknown and is represented by the prior $\underline{\theta}_0 \sim N(\underline{m}_0, \underline{C}_0)$ then the log-likelihood is (4.53) with

$$\mu^T = \underline{F}(\underline{G} \;\; \underline{G}^2 \ldots \underline{G}^n)\,\underline{m}_0$$

$$\Sigma_{ij} = \delta_{ij}V + \underline{F}\,\underline{G}^i\underline{C}_0\,(\underline{G}^T)^j\,\underline{F}^T + \underline{F}\sum_{k=1}^{i}\underline{G}^{i-k}\,\underline{W}\,(\underline{G}^T)^{j-k}\,\underline{F}^T.$$

Using the formulae $\dfrac{\partial}{\partial\phi_i}\log|\underline{\Sigma}| = \text{trace}\{\underline{\Sigma}^{-1}\dfrac{\partial\underline{\Sigma}}{\partial\phi_i}\}$

and $\frac{\partial \underline{\Sigma}^{-1}}{\partial \phi_i} = -\underline{\Sigma}^{-1} \frac{\partial \underline{\Sigma}}{\partial \phi_i} \underline{\Sigma}^{-1}$ in (4.53) gives the maximum likelihood estimates of ϕ_i - where the ϕ_i are the free parameters of $\{V, \underline{W}\}$ - as the solution of

$$\text{trace}\left\{\underline{\Sigma}^{-1} \frac{\partial \underline{\Sigma}}{\partial \phi_i}\right\} = (\underline{y}-\underline{\mu})^T \underline{\Sigma}^{-1} \frac{\partial \underline{\mu}}{\partial \phi_i} + \frac{\partial \underline{\mu}}{\partial \phi_i}^T \underline{\Sigma}^{-1} (\underline{y}-\underline{\mu}) + \\ + (\underline{y} - \underline{\mu})^T \left\{\underline{\Sigma}^{-1} \frac{\partial \underline{\Sigma}}{\partial \phi_i} \underline{\Sigma}^{-1}\right\} (\underline{y} - \underline{\mu}). \tag{4.54}$$

Apart from simple examples this is difficult to solve analytically and numerical methods are needed to find a solution. An added problem is the possible lack of identifiably, in that V and $\underline{W}$ might not be uniquely determined from the likelihood equation (4.54).

The difficulty of solving (4.54) implies similar obstacles to an exact Bayesian approach, since Bayesian estimation is essentially a modified likelihood method.

We shall illustrate some of the problems mentioned at the beginning of this section by considering the simplest practical DLM, namely the steady model of Example 4.4 defined in (4.31), (4.32) as

$$y_t = \theta_t + v_t \qquad v_t \sim N(0,V)$$
$$\theta_t = \theta_{t-1} + w_t \qquad w_t \sim N(0,W)$$

where V and W are now unknown.

The simplest method of estimating V and W consists of using the covariance properties of the 'derived' series $z_t = y_t - y_{t-1}$. (The justification for doing this is given later in Chapters 6 and 7). The properties are

$$\left.\begin{aligned} E(z_t) &= 0 \\ \gamma_0 &= E(z_t^2) = W + 2V \\ \gamma_1 &= E(z_t z_{t-1}) = -V. \end{aligned}\right\} \qquad (4.55)$$

Comparing these theoretical values with the estimated first two sample autocovariances c_0 and c_1 of the differenced data gives point estimates of V and W which can then be used in the Kalman filter.

This technique needs some data to start with, although it might be possible to use a weighted average of prior values and estimates to overcome this.

We shall develop very general results of equivalence between ARIMA models and DLMs in Chapters 6 and 7. It will be shown that an MA(1) model

$$z_t = y_t - y_{t-1} = \varepsilon_t + \beta\varepsilon_{t-1} \qquad (4.56)$$

is equivalent to the DLM (4.31), (4.32) where β really depends upon t. However in the limit as $t \to \infty$, from Chapter 6 or directly from Example 4.4

$$\beta = \frac{-W - 2V + \sqrt{(W^2+4VW)}}{2V} \qquad (4.57)$$

and the problem is to estimate β. This can be done by any classical approach, for example an asymptotic likelihood approach is presented in Example 2.4.

Having found β the ratio V/W is calculated from (4.57), which is all that is needed for filtering or forecasting since the Kalman filter for the steady model depends upon V, W only through the ratio V/W. If additionally individual values of V and W are needed then either the Bayesian technique described below can be used

or alternatively the variance σ^2 of (4.56) can be estimated and equated to V, W. For instance from (6.71) and Example 4.4

$$\sigma^2 = -V/\beta \quad . \tag{4.58}$$

Note that (4.57) and (4.58) follows from equating (4.55) to the autocovariances of (4.56). It will be proved in Chapter 6 that this is permissible for equivalence between (4.55) and (4.31), (4.32) using the Kalman filter as $t \to \infty$.

Exact Bayesian Analysis

The only case where 'nice' results can be derived appears to be when there is a single unknown parameter of such a form that the standard Normal-Gamma theory can be used. By 'nice' we mean that at each time stage priors and posteriors can be described in closed form corresponding to standard distributions. For example suppose that

$$y_t | y^{t-1}, \tau \sim N(m_t, 1/\tau)$$

that is the predictive distribution has unknown precision τ, and suppose further that conditional upon data up to time t-1, τ has a gamma distribution

$$p(\tau | y^{t-1}) \sim (\tfrac{1}{2}\beta)^{\frac{1}{2}\alpha} \frac{\tau^{\frac{1}{2}\alpha - 1} e^{-\frac{1}{2}\beta\tau}}{\Gamma(\frac{1}{2}\alpha)} \tag{4.59}$$

which we denote by $G[\frac{1}{2}\alpha, \frac{1}{2}\beta]$.

Then applying Bayes theorem

$$p(\tau | y^t) \propto p(y_t | y^{t-1}, \tau) p(\tau | y^{t-1}) \tag{4.60}$$

$$\sim G[\tfrac{1}{2}(\alpha+1), \tfrac{1}{2}\{\beta + (y-m)^2\}] \tag{4.61}$$

so that the posterior distribution is also a gamma distribution. The unconditional predictive distribution is

$$f(y_t|y^{t-1}) = \int \frac{(\beta/\alpha)}{\Gamma(\frac{1}{2}\alpha)\sqrt{(2\pi)}} e^{-\frac{1}{2}\tau\{\beta+(y-m)^2\}} \tau^{\frac{1}{2}(\alpha+1)-1} d\tau$$

$$\propto \{\beta + (y-m)^2\}^{-\frac{1}{2}(\alpha+1)}$$

which is a Student-t distribution.

These recursions have a simple form, however unfortunately the Kalman updating procedure involves two parameters $\underline{V}$ and $\underline{W}$ together with $\underline{\theta}_{t-1}|y^{t-1}$ and not a single precision τ. For example for the steady model (4.31), (4.32)

$$\tau_t = C_{t-1} + W + V$$

so that not only does τ_t provide information on the sum W+V but also C_{t-1} is involved additively.

These problems can be circumvented if we let $C_0=C_0*V$, $W=W*V$ with C_0*, $W*$ known so that there is just a single unknown V. The predictive distribution is then a constant multiplied by V, enabling the Normal-Gamma theory to be applied. Moreover the Kalman recursions depend only in $W/V = W*$, so that the analysis can be repeated at each stage. Unfortunately these assumptions are unlikely to be realistic precisely for these reasons, since knowledge of W/V completely determines the filter, as we have already mentioned. Equivalently from (4.55), the autocorrelations depend on W, V only through W/V.

It is interesting to note that if we put independent priors on V, W, both of which are inverse-gamma distributed

then we obtain an unconditional distribution for the system vector that is the first term in Smith's t-product (1979) which is mentioned in the next chapter. However again it is not clear how to proceed for the next time stage.

The problems of a fully Bayesian analysis can be seen by considering just one evolution for the steady model: if θ_0 is normally distributed then from the Kalman filter $p(\theta_1|y_1) \propto p(y_1|\theta_1,V)p(\theta_1|\theta_0)$ is also normally distributed if V and W are known. But if they are unknown, with priors p(V), p(W) say, then

$$p(\theta_1|y_1) = \int p(\theta_1|y_1,V,W)\ p(W,V|y_1)\ dVdW$$

which is a mixture of normal distributions weighted by

$$p(W,V|y_1) \propto p(y_1|V,W)\ p(V)\ p(W)$$

which removes all of the normality.

Approximate Bayesian Analysis

Instead of using a fully Bayesian analysis with priors for the unknowns and integrating at each time stage with respect to the unknowns to obtain the unconditional posteriors for the observation and system vectors, it is possible to approximate quantities. The simplest way of proceeding is to approximate the posteriors of the unknown parameters by singular distributions concentrated at a single point, so that the resulting integration consists of substituting a single value in the integrand. In other words point estimates of V and W are used at each time stage as inputs to the filter, the mean and variance

of the system vector are then updated via the recurrence relations, new posteriors for V and W are formed, point estimates derived and the process iterated. In general this approach involves discarding a lot of information.

For the steady model of Example 4.4, if $\theta_t|y_t \sim N(m_t, C_t)$ then the Kalman filter gives

$$\theta_{t+1}|y^t \sim N(m_t, C_t+W)$$
$$\theta_{t+1}|y^{t+1}, \tau \sim N(m_{t+1}, C_{t+1}) \qquad (4.62)$$

where

$$m_{t+1} = m_t + \frac{(W + C_t)}{(W+C_t+\tau^{-1})}(y_t - m_t) \qquad (4.63)$$

$$C_{t+1}^{-1} = (C_t + W)^{-1} + \tau \qquad (4.64)$$

where the precision $\tau \equiv V^{-1}$.

If the conditional distribution $p(\tau|y^t)$ has a gamma distribution (4.59), then using Bayes theorem (4.60) gives the posterior for τ as

$$p(\tau|y^{t+1}) \propto \frac{1}{(C_t+W+\tau^{-1})^{\frac{1}{2}}} \exp\left\{\frac{-\frac{1}{2}(y-m_t)^2}{C_t+W+\tau^{-1}} - \tfrac{1}{2}\beta\lambda\right\} \tau^{\frac{1}{2}\alpha-1} \qquad (4.65)$$

If in addition W is known - so that τ is the only unknown - then the exact Bayesian procedure is to find the posterior of $\theta_{t+1}|y^{t+1}$ by integrating (4.62) with respect to τ having the measure (4.65). However the resultant distribution will not then be normal. An approximate method is to derive a point estimate of τ from (4.65) (or less satisfactorily from $p(\tau|y^t)$) and then substitute this into (4.63) - (4.64). The posterior $p(\theta_{t+1}|y^{t+1})$ will then be normal with parameters m_{t+1}, C_{t+1}. Unfortunately

(4.65) does not have a particularly convenient form, but if we approximate this by a gamma distribution then the procedure can be repeated iteratively.

This leaves the problem of how to best approximate (4.65) by a gamma distribution. West (1981) considers general symmetric error terms for v_t, and in the normal case effectively uses

$$p(\tau|y^{t+1}) \propto \tau^{\frac{1}{2}(\alpha+1)-1} e^{-\frac{1}{2}\tau\{\beta+(y_t-m_t)^2\}}.$$

At each time stage the posterior is then gamma distributed

$$p(\tau|y^t) = G[\tfrac{1}{2}\alpha_t, \tfrac{1}{2}\beta_t]$$

with the recursions

$$\alpha_t = \alpha_{t-1} + 1$$

$$\beta_t = \beta_{t-1} + (y_{t-1} - m_{t-1})^2.$$

The mode or the mean $(\alpha_{i+1} / \beta_{i+1})$ can then be used as the point estimate of τ in (4.63), (4.64).

This is tantamount to approximating $(C_t+W+\tau^{-1})$ by τ^{-1} which is not without its disadvantages. For example (4.65) can become multimodal whereas it is being approximated by a unimodal distribution. The multimodality can be demonstrated by differentiating (4.65) with respect to τ. Equating this derivative to zero gives a cubic in τ, and if for instance $|y-m_t|$ is large enough it is straight - forward to show that the cubic has three positive roots, thereby demonstrating multimodality.

CHAPTER 5

NON-NORMAL MODELS

5.1 Introduction

Chapter 4 reviewed Bayesian Forecasting which is a theory of forecasting based upon a 'state-space' description of a time series, Bayesian inference and Normal error distributions; the Kalman Filter then enables us to recursively update our beliefs as expressed in the appropriate posterior and predictive distributions. In this chapter we look briefly at the problems encountered when non-Normal error distributions are introduced and when, more generally, the state and observation equations are redefined.

There is currently much interest in both non-stationary and non-linear Time Series, for example in the papers of Priestley (1980), Haggan and Ozaki (1981), Lawrance and Lewis (1980) and the piece-wise linear work of Tong in, for example, Tong and Lim (1980). Our attention will be focussed more on non-linear forecasting models or non-linear forecasts. Indeed it is possible for a linear model to yield non-linear forecasts, that is the forecasts are non-linear functions of the data; for instance the application of the Class 11 procedure to a time-invariant DLM gives posterior means and variances which are not linear in the observations - since we are effectively updating the mixture proportions - consequently the forecasts obtained from the predictive distributions will in general be non-linear. This example highlights the importance of loss functions in forecasting, because the predictive distributions are Normal mixtures

so that any point forecast will be heavily dependent on the loss function used.

5.2 General Filtering Theory

The state-space representation introduced in Chapters 3 and 4 consists of an observation equation relating the observations to an underlying system parameter $\underline{\theta}_t$ which undergoes a Markov evolution. In discrete time the system process is a Markov chain on a continuous vector space Θ which we shall assume is a subspace of $\mathbb{R}^n$. We shall assume further that this process can be described by a transition density with respect to Lebesgue measure on $\mathbb{R}^n$, and that for each $\theta\varepsilon\Theta$ there is a probability density $f(\underline{x}|\underline{\theta})$ with respect to a fixed measure μ on the observation space X, the latter being a subset of the real line or more generally a subset of $\mathbb{R}^p$. In practice μ will either be Lebesgue measure or such that $f(\underline{x}|\underline{\theta})$ represents a discrete density function.

Suppose that the process starts at time t=0 and that the initial state is described by a density $p(\underline{\theta}_0)$, then at time t the observation has density function

$$p(\underline{x}_t) = \int f(\underline{x}_t|\underline{\theta}_t)p(\underline{\theta}_t)\, d\theta_t \qquad (5.1)$$

where $p(\underline{\theta}_t)$ is the density function of the observation vector at time t, given by

$$p(\underline{\theta}_t) = \int\cdots\int p(\underline{\theta}_t|\underline{\theta}_{t-1})p(\underline{\theta}_{t-1}|\underline{\theta}_{t-2})\cdots p(\underline{\theta}_1|\underline{\theta}_0)p(\underline{\theta}_0)d\theta_0\cdots d\underline{\theta}_{t-1}. \qquad (5.2)$$

The simplest such models are those in which $p(\underline{\theta}_t|\underline{\theta}_{t-1})$ and $f(\underline{x}_t|\underline{\theta}_t)$ do not depend on t.

By placing certain restrictions on the unconditional densities (5.1) and (5.2) we can obtain conditions on the pair $\{f(\underline{x}|\underline{\theta}), p(\underline{\phi}|\underline{\theta})\}$, for example we might require (5.1) to be of the same type for all t. However we are more interested in conditional inference, or sequential, although (5.1) and (5.2) should have a 'nice' form for all t if we are to be able to simulate the process.

Suppose that $p(\underline{\theta}_0)$ is the prior distribution for $\underline{\theta}$ at time t=0, then the posterior at time t is given by

$$p(\underline{\theta}_t|\underline{y}^t) = \frac{f(\underline{y}_t|\underline{\theta}_t)p(\underline{\theta}_t|\underline{y}^{t-1})}{\int f(\underline{y}_t|\underline{\theta}_t)p(\underline{\theta}_t|\underline{y}^{t-1})d\theta_t} \tag{5.3}$$

where $$p(\underline{\theta}_t|\underline{y}^{t-1}) = \int p(\underline{\theta}_t|\underline{\theta}_{t-1})p(\underline{\theta}_{t-1}|\underline{y}^{t-1})d\underline{\theta}_{t-1}. \tag{5.4}$$

The relevant predictive densities are obtained from

$$p(\underline{\theta}_{t+k}|\underline{y}^t) = \int p(\underline{\theta}_{t+k}|\underline{\theta}_{t+k-1})p(\underline{\theta}_{t+k-1}|\underline{y}^t)d\underline{\theta}_{t+k-1} \tag{5.5}$$

and

$$f(\underline{y}_{t+k}|\underline{y}^t) = \int f(\underline{y}_{t+k}|\underline{\theta}_{t+k})p(\underline{\theta}_{t+k}|\underline{y}^t)d\underline{\theta}_{t+k}\ . \tag{5.6}$$

As we have seen in Chapter 4, under assumptions of normality all the equations (5.1) - (5.6) possess a particularly simple form.

If the predictive density functions of the observations are to be evaluated easily we require that the densities obtained from (5.6) via (5.3) - (5.5) should be tractable for each k and t. We can write from above

$$f(\underline{y}_{t+k}|\underline{y}^t) =$$
$$\int f(\underline{y}_{t+k}|\underline{\theta}_{t+k})\int\cdots\int p(\underline{\theta}_{t+k}|\underline{\theta}_{t+k-1})\cdots p(\underline{\theta}_{t+1}|\underline{\theta}_t)p(\underline{\theta}_t|\underline{y}^t)d\underline{\theta}_{t+k}\cdots d\underline{\theta}_t \tag{5.7}$$

which will be 'tractable' if for no k or t do we have to resort

to numerical methods for its calculation, which will be the case if the densities (5.7) are 'standard' density functions or can be expressed analytically in closed form.

Three problems associated with using non-normal recursions are

(1) Finding pairs $\{f(\underline{x}|\underline{\theta}), p(\underline{\phi}|\underline{\theta})\}$ which lead to tractable results in the above sense.

(2) Identifying subsets of the admissible pairs in (1) which do not need the past history $\underline{y}_1 \ldots \underline{y}_t$ stored for each t. More precisely Bather (1965) requires that there exists a sequence of statistics $u_t(y_1 \ldots y_t)$ such that $p(\underline{\theta}_t|\underline{y}^t)$ {and hence all the other predictive distributions} depends on $\underline{y}^t$ only through $\underline{u}_t$. This is a stochastic type of sufficiency.

(3) Interpreting such models - that is choosing models of the form satisfying (1) and (2) which correspond to a reasonable model for the situation under consideration, rather than being just of a convenient mathematical form.

Example 5.1

Consider the steady model of Chapter 4

$$y_t = \theta_t + v_t \tag{5.8}$$

$$\theta_t = \theta_{t-1} + w_t . \tag{5.9}$$

This widely used model has a ready interpretation, with v_t and w_t error terms. Usually these are taken to be Normal, with mean 0 and variance V, however we can think of situations where it might be preferable to have non-normal error terms. For example following the work of Huber (1981) we might want to put a heavy tailed distribution on v_t or

w_t to 'protect' us against outliers; this is very closely linked to the 'outlier resistant' distributions of O'Hagan (1979) and has motivated West's (1981) work.

The next two examples use (5.8) and (5.9) with non-normal errors.

Example 5.2

Let w_t have a strictly Levy-stable distribution of characteristic exponent α independent of t, W say, and let v_t be strictly Levy-stable of the same type; definitions are given in say Feller (1970, page 170). This gives a general class of steady models which can be simulated.

From (5.9) assuming that $\theta_0=\mu$ with probability one (we could equally easily assume a Levy-stable distribution with location μ, exponent α),

$$\theta_T = \mu + w_1 \ldots + w_T$$
$$\sim \mu + T^{1/\alpha} W$$

by definition of stable distributions. By assumption, $v_T = CW$ for some positive C, so

$$y_T = \theta_T + v_T$$
$$\sim (T + C^{\alpha})^{1/\alpha} W + \mu \qquad (5.10)$$

using the fact that $s^{1/\alpha} x_1 + t^{1/\alpha} x_2 \sim (s + t)^{1/\alpha} x$ for all strictly stable laws x, with x_1, $x_2 \sim x$, which holds true for all $s,t > 0$. Thus each marginal predictive distribution (5.10) is Levy-stable with exponent α.

The above assumptions can also be used with the more general steady model introduced in Chapter 7 which replaces (5.8) by

$$y_t = \theta_t + \theta_{t-1} + v_t$$

giving

$$y_T = 2\mu + 2(w_1 \ldots + w_{T-1}) + w_T + v_T$$

$$\sim 2\mu + (2^{\alpha}(T-1) + 1 + C^{\alpha})^{1/\alpha} W$$

in place of (5.10).

The limitations of this class of models lies in the complicated nature of the Levy-stable density functions. The case $\alpha = 2$ corresponds to the Normal distribution and the results of Chapter 4; with $\alpha = 1$, we have the Cauchy distribution giving the following .

Example 5.3

Consider the case where v_t, w_t have a Cauchy distribution, which has form

$$f(x) = \frac{1}{\pi} \frac{t}{t+(x-\mu)^2}$$

where t is a scale parameter and μ the location. If we suppose that v_t is Cauchy with location zero, scale v, w_t Cauchy location zero, scale w then under the analysis of Example 5.2

y_T is a Cauchy distribution with location μ and scale $\eta w+v$.

This example illustrates the remarks made earlier, in that the posterior distributions of (5.3) and (5.4) have a complicated form. Moreover if we have

$$\theta_t = \theta_t + w_t$$

with w_t a Cauchy distribution and suppose that $\theta_{t-1}|y^t$ has a Cauchy distribution, then it is not possible to find $f(y|\theta)$ such that $p(\theta_t|y_t)$ has a Cauchy distribution for all t.

5.3 Predictive Consequences of Non-Gaussian Steady Models

The last section contained examples where the convolution of densities lead to tractable results, because the relevant distributions were stable. If additive error terms are used it is difficult to work outside such a class of distributions, otherwise approximations have to be made as in West (1981) for example.

We now consider alternative specifications of state-space models, and consider the predictive implications of such models. That is we discuss the predictive distributions and point estimates under various loss functions which correspond to the classical k-step ahead predictors. In particular we consider the exponential family and show that for models which undergo a suitably defined evolution, non-normal distributions can give rise to behaviour very different from the classical steady model. For example the condition (4.33)

$$y_t(\ell) = y_t(1)$$

which says that the forecast are constant at all lead times ℓ need no longer hold. The results given amplify the initial work of Key and Godolphin (1981).

From now on we only consider models whose system parameter is univariate with observation space a subset of the real line.

An alternative to specifying the system evolution by a transition density $p(\theta_t|\theta_{t-1})$ is to define the conditional evolution of $\theta_{t-1}|y^t$ to $\theta_t|y^t$. The quantities of interest can be calculated from (5.3) - (5.6), and if a 'sensible' evolution is defined some of the integration problems can be avoided.

This is the approach of Smith (1979), motivated by the steady model of Example 4.4 expressed in (5.8) and (5.9) with normal errors. Such a model can be rewritten as

$$y_t|\theta_t \sim N(\theta_t, V) \tag{5.11}$$

with the evolution

$$p(\theta_{t+1}|y^t) \propto \{p(\theta_t|y^t)\}^{k_t} \tag{5.12}$$

where $k_t = C_t/(C_t + W)$, $W>0$. As t increases the prior effects disappear, $C_t \to C$, $k_t \to k$, with $0<k<1$, where k depends on V and W. Then

$$m_t = E[\theta_t|y^t]$$

is obtained from

$$m_t = m_{t-1} + (1-k)(y_t - m_{t-1}) \tag{5.13}$$

so that

$$m_t = (1-k) \sum_{j=0}^{\infty} k^j y_{t-j} \tag{5.14}$$

the familiar EWMA of (2.15) and (4.35).

Using this example, Smith abstracts two requirements for his models:

(i) Decisions about θ_t at times t and t+1 conditional upon information up to time t should be the same

(ii) The uncertainty associated with such decisions should increase.

As remarked upon in Chapter 4, decisions depend on the choice of loss function, so that the above requirements generate a whole class of decision based steady models, which are 'subjective' in that they depend upon the loss function used. To obviate this problem, Smith restricts his attention to his so-called 'utility-invariant' loss functions, namely the step-loss functions of gauge b defined by

$$L_b(y-d) = \begin{cases} 0 & |y-d| \leq b \quad b>0 \\ 1 & \text{otherwise} \end{cases} \tag{5.15}$$

Using such loss functions with (i) and (ii), together with an additional requirement to make the results independent of any preconditioning of θ_t to lie in a specific interval, Smith arrives at his 'steady model'. This is a model described by a probability density for $y_t|\theta_t$ for which (5.12) holds with k_t independent of t,

$$f(y_t|\theta_t) \tag{5.16}$$

$$p(\theta_{t+1}|y^t) \propto \{p(\theta_t|y^t)\}^k. \tag{5.17}$$

In fact such a 'steady evolution' can always be expressed via a transition density $p(\theta_t|\theta_{t-1})$ through (5.4). This is because subject to certain regularity requirements, the Radon-Nikodym theorem states that there is a function $p(\theta_t|\theta_{t-1})$ such that (5.4) holds. However, in general such a function will depend on t and may not have a closed form.

In a forecasting context the main quantities of interest are the joint and marginal predictive densities, $f(y_{t+k} \cdots y_{t+1}|y^t)$ and $f(y_{t+k}|y^t)$. Although Smith's formulation appears to be only a one-step ahead phenomenum, in fact (5.16) and (5.17) enable us to calculate all the required predictive densities. For example (5.6) gives $f(y_{t+1}|y^t)$ while

$$f(y_{t+2}|y^t) = \int f(y_{t+2}, y_{t+1}|y^t)dy_{t+1}$$
$$\int f(y_{t+2}|y^{t+1})f(y_{t+1}|y^t)dy_{t+1}.$$

The joint predictive densities are given by

$$f(y_{t+\ell}, \cdots y_{t+1}|y^t) = f(y_{t+\ell}|y^{t+\ell-1})f(y_{t+\ell-1}|y^{t+\ell-2}) \cdots f(y_{t+1}|y^t) \tag{5.18}$$

and the marginal densities $f(y_{t+\ell}|y^t)$ are obtained from (5.18) by integrating out $y_{t+\ell-1} \cdots y_{t+1}$.

Example 5.4

In the normal steady model of Example 4.4 or above

$$y_{t+1}|y^t \sim N(m_t, C_t+W+V)$$

and in the limit $C_t \to C$, but $k=C/C+W$ or $C=W(k/1-k)$ so

$$y_{t+1}|y^t \sim N(m_t, \frac{W}{1-k} + V)$$

and

$$y_{t+\ell}|y^t \sim N(m_t, C_t + \ell W + V)$$

$$= N(m_t, W[\frac{k+\ell(1-k)}{1-k}] + V).$$

Consequently using any loss function symmetric in $\{y_t(\ell)-y_{t+\ell}\}$ we have

$$y_t(\ell) = y_t(1) = m_t \qquad (5.19)$$

so that using (5.13)

$$y_t(\ell) = y_{t-1}(\ell+1) + (1-k)\{y_t - y_{t-1}(1)\} \qquad (5.20)$$

where m_t is the EWMA (5.14).

Equations (5.19) and (5.20) are the familiar defining relations for a steady forecasting model and for a predictor updating equation respectively. These results are detailed in Box and Jenkins (1970, chapter 5) and have been described under a variety of assumptions by Holt (1957), Brown (1959), Muth (1960) and Whittle (1963, chapter 8). The steady forecast equation (5.19) suggests that the model is trend free, a condition which is independent of any model assumptions and a point discussed in Godolphin and Harrison (1975). Equations (5.19) and (5.20) imply that $y_t(\ell)$ is

the EWMA (5.14), however the 'discount factor' k assumes only the positive part of its natural range $|k|<1$. This point is developed further in Chapters 6 and 7.

The two equations (5.19) and (5.20) are a more natural example of a steady forecasting model than the equations giving a steady evolution of the system parameter. However the latter based upon (5.16) and (5.17) are well defined and it is interesting to examine the consequences of these assumptions in terms of predictions.

To obtain general results we shall consider the exponential family of distributions, so that the observation equation is of the form

$$f(y_t|\theta_t) = \exp\{a(\theta_t)b(y_t) + c(\theta_t) + d(y_t)\}. \qquad (5.21)$$

All of the examples in Smith (1979) lie in this family apart from the interesting Student-T steady model. Under requirement (2) of §5.2 and certain differentiability conditions Bather (1965) shows that under a given Markov evolution $p(\theta_t|\theta_{t-1}), f(y_t|\theta_t)$ must be of the form (5.21). Strictly speaking we need to apply the mild restrictions that the distributions (5.23) are strictly identifiable with respect to the dominating measure and also that the sample space Y does not depend upon θ, that is $f(y|\theta)>0$ always.

In order to satisfy requirement (1) of §5.2 we shall assume that at each time stage the posterior distribution $p(\theta_t|y^t)$ has the conjugate form

$$p(\theta_t|y^t) \propto \exp\{\gamma_t a(\theta_t) + \delta_t c(\theta_t)\} \qquad (5.22)$$

for some parameters γ_t, δ_t. Then under (5.17)

$$p(\theta_{t+1}|y^t) \propto \exp\left[k\{\gamma_t a(\theta_{t+1}) + \delta_t c(\theta_{t+1})\}\right]. \qquad (5.23)$$

Consequently using (5.21) and (5.3) we have the recurrence relations

$$\gamma_{t+1} = k\gamma_t + b(y_{t+1}) \ , \ \delta_{t+1} = k\delta_t + 1 \qquad (5.24)$$

which for sufficiently large t have the solution

$$\gamma_t \underset{\sim}{=} u_t = \sum_{j=0}^{\infty} k^j b(y_{t-j}) \ , \ \delta_t \underset{\sim}{=} (1-k)^{-1}. \qquad (5.25)$$

In the limit with δ_t given by (5.25), (5.22) and (5.23) form the 'invariant conditional densities' of Bather (1965), that is both densities depend upon $y_1 \ldots y_t$ only through γ_t, which is the appropriate sequence of real valued functions mentioned in criterion 2 of §5.2.

If we substitute the limiting values of (5.25) into (5.22), (5.23) and define the functions γ, ρ by

$$\exp\{\lambda(u_t)\} = \int \exp\{a(\theta)u_t + \frac{c}{1-k}(\theta)\}\, d\theta \qquad (5.26)$$

$$\exp\{\rho(ku_t)\} = \int \exp\{ka(\theta)u_t + \frac{k}{1-k}\, c(\theta)\} d\theta \qquad (5.27)$$

where the definitions hold for those regions for which the integrals are finite, then the predictive distribution (5.6) is given by

$$\begin{aligned} f(y_{t+1}|y^t) &= \int \exp\{a(\theta_{t+1})b(y_{t+1}) + c(\theta_{t+1}) + d(y_{t+1}) + ka(\theta_{t+1})u_t \\ &\qquad + \frac{k}{1-k} c(\theta_{t+1}) - \rho(ku_t)\} d\theta_{t+1} \\ &= \exp\{\lambda(u_{t+1}) - \rho(ku_t) + d(y_{t+1})\} \qquad (5.28) \end{aligned}$$

with $\quad u_{t+1} = b(y_{t+1}) + ku_t = \sum_{j=0}^{\infty} k^j b(y_{t+1-j}).$

The joint predictive distributions are given by

$$f(y_{t+\ell}, y_{t+\ell-1}, \cdots y_{t+1}|y^t) = \exp\{\sum_{i=1}^{\ell} \lambda(u_{t+i}) - \rho(ku_{t+i-1}) + d(y_{t+i})\}, \qquad (5.29)$$

and the marginal predictive densities can be calculated by integration, for example

$$f(y_{t+2}|y^t) = \exp\{\alpha(y_{t+2}) - \rho(ku_t) + d(y_{t+2})\} \qquad (5.30)$$

where

$$\exp\{\alpha(y_{t+2})\} = \int \exp[\lambda\{b(y_{t+2}) + ku_{t+1}\} + \lambda(u_{t+1}) - \rho(ku_{t+1}) + d(y_{t+1})]\, d\upsilon(y_{t+1}) \qquad (5.31)$$

where $\upsilon(y)$ in the appropriate measure, so that we include discrete distributions.

Such a system gives the invariant conditional distributions defined by Bather (1965) provided that the system evolution corresponds to a Markov evolution, that is provided that there is a function $p(\phi|\theta,u)$ satisfying

$$\exp\{kua(\phi) + \frac{k}{1-k} c(\phi) - \rho(ku)\} = \int p(\phi|\theta) \exp\{ua(\theta) + \frac{1}{1-k} c(\theta) - \lambda(u)\} d\theta \qquad (5.32)$$

such that

(i) $p(\phi|\theta,u)$ is a non-trivial function of θ [so that $\frac{\partial}{\partial\theta} p(\phi|\theta,u) \neq 0$] and for each $\theta_t \varepsilon \Theta$ $p(\phi|\theta,u)$ is a probability function.

(ii) $p(\phi|\theta,u)$ is independent of u

(iii) $p(\phi|\theta,u)$ is integrable with respect to the dominating measure μ.

If these conditions are satisfied, then $p(\phi|\theta)$ will be a (stationary) transition probability, since for example we can always satisfy (5.32) with

$$p(\phi|\theta,u) = \exp\{kua(\phi) + \frac{k}{1-k} c(\phi) - \rho(ku)\}.$$

We can show that

Theorem 5.5

A transition density $p(\phi|\theta)$ exists which satisfies the conditions (i), (ii), (iii) and (5.32) if and only if there is a function v such that

$$\int v(\psi)\exp(u\psi)d\psi = \exp\{\lambda(u) - \rho(ku)\} \qquad (5.33)$$

the equation holding for the values of u such that the right-hand side is well defined.

Proof

The necessity of the result follows from the work of Bather (1965) with only minor modifications, these are needed because we have a function $a(\theta)$ instead of the function θ. We briefly give the mainsteps in the proof with the required modifications. Here ϕ denotes the system parameter at the next time stage from that for θ, at say t+1 and t respectively. If we reverse the natural order and look at θ conditional upon ϕ, then if a transition density exists

$$E\{e^{i\zeta a(\theta)}|\phi,u\} = \int p(\phi|\theta)\frac{p(\theta|u)}{p(\phi|u)}e^{i\zeta a(\theta)}\,d\theta$$

which using (5.22) and (5.27) using the limiting values (5.25)

$$= \int p(\phi|\theta)\frac{\exp\{a(\theta)(u+i\zeta) + (1-k)^{-1}c(\theta) - \lambda(u)\}d\theta}{\exp\{kua(\phi) + k(1-k)^{-1}c(\phi) - \rho(ku)\}}$$

which using (5.32)

$$= \exp\left[\, a(\phi)ki\zeta + \lambda(u+i\zeta)-\lambda(u)+\rho(ku)-\rho\{k(u+i\zeta)\}\,\right].$$

This implies that $E[e^{i\zeta\{a(\theta) - ka(\phi)\}}|\phi,u]$ does not depend on ϕ and so ϕ and $a(\theta)-ka(\phi)$ are independent random variables.

Putting $\psi=a(\theta)-ka(\phi)$ then the two alternative forms for the joint pdf of ϕ and θ are

$$p(\phi|\theta)\exp\{a(\theta)u + \frac{c(\theta)}{1-k} - \lambda(u)\}$$

and $\exp\{a(\phi)ku + \frac{k}{1-k}c(\phi) - \rho(ku)\}|a'(\theta)|\omega\{a(\theta) - ka(\phi)|u\}$

where ψ has the conditional density $\omega(\psi|u)$. In the above we have used the fact that $a(\theta)=\psi+ka(\phi)$, so that after transforming variables the density of θ conditional on ϕ and u is $|a'(\theta)|\omega\{a(\theta)-ka(\phi)|u\}$, provided that the function a is 1-1. On replacing $a(\theta)$ by $\psi+ka(\phi)$ in the above we have

$$\omega(\psi|u) = p\left[\phi|a^{-1}\{\psi+ka(\phi)\}\right]\exp\{\psi u+\rho(ku)-\lambda(u)\}\ x$$

$$\exp\left(\frac{c\left[a^{-1}\{\psi+ka(\phi)\}\right]}{1-k} - \frac{kc(\phi)}{1-k}\right)\ |a'\{\psi+ka(\phi)\}|^{-1}$$

which is of exponential type and so can be written

$$\omega(\psi|u) = v(\psi)\exp\{\psi u-\lambda(u)+\rho(ku)\}$$

so that $v(\psi)$ satisfies $\int v(\psi)\exp(\psi u)d\psi = \exp\{\lambda(u)-\rho(ku)\}$ as required. This proves necessity - in fact we have also proved the existence of ψ in this case. Conversely suppose that such a function exists and define

$$p(\phi|\theta) = \exp\{a(\phi)ku - a(\theta)u + \frac{k}{1-k}c(\phi) - \frac{c(\theta)}{1-k} + \lambda(u) - \rho(ku)\}\ x$$

$$|a'(\theta)|v\{a(\theta) - ka(\phi)\}\exp\left[\{a(\theta) - ka(\phi)\}u - \lambda(u) + \rho(ku)\right]$$

$$= \exp\left\{\frac{k}{1-k}c(\phi) - \frac{c(\theta)}{1-k}\right\}|a'(\theta)|v\{a(\theta) - ka(\phi)\}.$$

Then $$\int p(\phi|\theta)\exp\{a(\theta)u + \frac{c(\theta)}{1-k} - \lambda(u)\}d\theta$$

$$= \exp\{\frac{k}{1-k}c(\phi) - \lambda(u)\}\int e^{a(\theta)u}|a'(\theta)|\{a(\theta) - ka(\phi)\}d\theta,$$

which putting $\psi = a(\theta) - ka(\phi)$

$$= \exp\{\frac{k}{1-k} c(\phi) - \lambda(u) + ka(\phi)u\} \int e^{\psi u} v(\psi) d\psi$$

$$= \exp\{uka(\phi) + \frac{k}{1-k} c(\phi) - \rho(ku)\}$$

so that (5.32) is satisfied and the result is proved.

This theorem enables us to reduce the steady models to the form

$$f(y_t|\theta_t)$$

$$p(\theta_t|\theta_{t+1})$$

provided that we can solve the transform equation (5.33).

Example 5.6

If we consider the Gamma-Gamma steady model of Example 5.8 then

$$\exp\{\lambda(u) - \rho(ku)\} = cu^{-1} \quad \text{where c is a constant.}$$

But $$\int_0^\infty e^{-st} dt = \frac{1}{s}$$

so that $v(\psi) = \begin{cases} 0 & \psi > 0 \\ c & \psi \leq 0 \end{cases}$

satisfies the required integral equation and so since for this example $a(\theta) = -\theta$, $c(\theta) = \alpha \log \theta$

$$p(\phi|\theta) = c\left(\frac{\phi}{\theta}\right)^{k/1-k} \frac{1}{\theta} \quad , \quad 0 < k\phi < \theta$$

which is a Beta distribution. Note that in this example $(-\psi)$ has an exponential distribution.

If equations (5.21) - (5.23) define a steady model in the sense that the predictors are trend free then we expect (5.19) to be satisfied for $\ell \geq 2$. It would appear however from (5.28) and (5.30) that even in the case $\ell = 2$ we would often not obtain $y_t(2) = y_t(1)$ except possibly if the functional form of the loss function depends upon the shape

of the distribution. For example there are several members of the discrete form of the exponential family where the expectations differ, that is $E[y_{t+2}|y^t] \neq E[y_{t+1}|y^t]$ corresponding to the use of quadratic loss functions. We are once again in a situation where our forecasts are dependent upon the choice of loss function. There are two possible cases where we can obtain some measure of independence from the loss function:

Case 1 Where the marginal predictive distributions are symmetric and unimodal since then any symmetric loss function yields the mean (which equals the mode) as the point estimate.

Case 2 Where the predictive distributions undergo a steady evolution as defined by Smith, that is

$$f(y_{t+2}|y^t) \propto T\{f(y_{t+1}|y^t)\}$$

for a convex function T and in particular

$$f(y_{t+2}|y^t) \propto \{f(y_{t+1}|y^t)\}^{k_0} \qquad (5.34)$$

for some k_0, $0<k_0<1$. There are several well defined members of the exponential family steady models where (5.34) is not satisfied. A similar set of remarks may be made concerning the Kalman updating equation (5.20). It follows that the examples satisfying Smith's definition of a steady model do not always satisfy the familiar expressions for predictors of the steady model given by the EWMA.

On the other hand Smith derives the EWMA in many of his particular models and we can generalise his argument as follows. Take the limiting forms for δ_t, γ_t defined by (5.25) and let $\hat{\theta}_t$ denote the mode of $\theta_{t+1}|y^t$ which from (5.17) is

also the mode of $\theta_t|y^t$. Then θ_t satisfies

$$\gamma_t a'(\hat{\theta}_t) + \delta_t c'(\hat{\theta}_t) = 0 \qquad (5.35)$$

provided that the supremum of (5.2) does not occur at a boundary point of the support and that a(.) and c(.) are differentiable. But $f(y_t|\theta_t)$ is a probability distribution and so

$$\int \exp\{a(\theta_t)b(y_t) + c(\theta_t) + d(y_t)\}\, d\nu(y_t) = 1 \qquad (5.36)$$

where $\nu(y_t)$ is either Lebesgue measure or a discrete measure. Provided that sufficient regularity exists,(which is always true for discrete measures otherwise we require $|a'(\theta)b(y) + c'(\theta)|\exp\{a(\theta)b(y) + c(\theta) + d(y)\} \leq g(y)$ for an integrable function g(y) for all y,θ which will be satisfied if a'(θ) and c'(θ) are bounded and if the expectation of b(y) with respect to f(y|θ) exists for all θ), then we can differentiate (5.36) to obtain

$$\int \{a'(\theta)b(y) + c(\theta)\ f(y|\theta) = 0$$

that is $\quad a'(\theta)E[b(y_t|\theta_t)] + c'(\theta_t) = 0$

and so for sufficiently large t

$$E[b(y_{t+1}|\theta_t)]_{\theta_t=\hat{\theta}_t} = \gamma_t\delta_t^{-1} \sim (1-k)\sum_{j=0}^{\infty} k^j b(y_{t-j}) \qquad (5.37)$$

where $\hat{\theta}_t$ is a mode. So that under the natural parametrisation of the exponential family with $b(y_t)\equiv y_t$ the expectation $E[y_{t+1}|\theta_t]|_{\theta_t=\hat{\theta}_t}$ is given by an EWMA. In general the quantity on the left hand side of (5.37) would not be considered a sensible forecast since to use it would be to discard much of the information about θ_{t+1} contained in $p(\theta_{t+1}|y^t)$.

Finally it is perhaps worth pointing out that the

steady evolution described by (5.18) is essentially a one-step phenomenum. This is because

$$p(\theta_{t+2}|y_t) = \int p(\theta_{t+2}|y^{t+1}) f(y_{t+1}|y^t) dy_{t+1}$$

$$= \int \exp\{ku_{t+1}a(\theta_{t+2}) + \frac{k}{1-k}c(\theta_{t+2}) - \rho(ku_{t+1})$$

$$+\lambda(u_{t+1}) - \rho(ku_t) + d(y_{t+1})\, dy_{t+1}.$$

But $$u_{t+1} = ku_t + b(y_{t+1})$$

so

$$p(\theta_{t+2}|y^t) = \exp\{k^2 u_t a(\theta_{t+2}) + \frac{k}{1-k}c(\theta_{t+2}) - \rho(ku_t) + \beta(\theta_{t+2})\} \tag{5.38}$$

where

$$\exp\{\beta(\theta_{t+2})\} = \int \exp\{a(\theta_{t+2})kb(y_{t+1}) - \rho(ku_{t+1}) + \lambda(u_{t+1}) + d(y_{t+1})\} dy_{t+1} \tag{5.39}$$

so that in general $p(\theta_{t+2}|y^t) \propto p(\theta_{t+1}|y^t)^{k_0}$, for any k_0, $0<k_0<1$, the latter density given by (5.23), (5.25) as

$$\left[\exp\{ku_t a(\theta_{t+1}) + \frac{k}{1-k}c(\theta_{t+1}) - \rho(ku_t)\right]^{k_0}.$$

5.4 Examples of Non-Normal Evolutions

This section illustrates the points raised in the previous sections by giving particular examples of the exponential family. We shall only consider the limiting, steady-state forms to avoid results being dependent upon assumptions concerning the prior.

First Example 5.7 considers the normal case and shows that the typical features of the EWMA forecast function follow when the loss function is symmetric. Example 5.8 considers the Gamma-exponential model. In this case it is shown that the forecast function depends heavily on the

loss function, for example under step loss functions the forecast function is a constant which is independent of the data, and for which the uncertainty associated with the forecasts can decrease. However under quadratic loss this model has many of the features of the standard normal model. Similar remarks apply to the Gamma-Gamma model of Example 5.13 which extends Example 5.8. Lastly Example 5.14, the Beta-Binomial shows that it is possible to have neither of the EWMA criteria (5.19), (5.20) satisfied when we have non-normal recursions.

Example 5.7

For comparison purposes we show how the Harrison-Stevens model of Examples 5.1 and 5.4 falls into the exponential form. The observations are normal

$$y_t|\theta_t \sim N(\theta_t, V)$$

which corresponds to (5.21) with

$$a(\theta) = \theta/V$$
$$b(y) = y$$
$$c(\theta) = -\theta^2/2V$$
$$d(y) = \frac{-y^2}{2V} - \tfrac{1}{2}\ln(2\pi v).$$

Equations (5.22) - (5.25) give

$$\theta_t|y^t \sim N(m_t, (1-k)V)$$
$$\theta_{t+1}|y^t \sim N(m_t, (1-k)V/k)$$

where

$$u_{t+1} = ku_t + y_{t+1}, \quad m_t = (1-k)u_t$$

$$e^{\lambda(u_t)} = \sqrt{2\pi(1-k)V}\exp\{u^2\frac{(1-k)}{2V}\}$$

$$e^{\rho(ku_t)} = \sqrt{2\pi\frac{(1-k)}{k}V}\ \exp\left\{\frac{ku^2(1-k)}{2V}\right\}.$$

The predictive distributions (5.28), (5.30) are

$$y_{t+1}|y^t \sim N(m_t, V/k) \tag{5.40}$$

$$y_{t+2}|y^t \sim N(m_t, \frac{V}{k}\{(1-k)^2+1\}) \tag{5.41}$$

and the joint predictive distributions (5.29)

$$f(y_{t+\ell}\cdots y_{t+1}|y^t) = \frac{1}{(2\pi V/k)^{\ell/2}}\exp\left[\frac{1}{2V}\sum_{i=1}^{\ell}\{-y_{t+i}^2+u_{t+i}^2(1-k)-k(1-k)u_{t+i-1}^2\}\right]$$

$$= \frac{1}{(2\pi V/k)^{\ell/2}}\exp\left[\frac{-k}{2V}\sum_{i=1}^{\ell}\{y_{t+i}-(1-k)u_{t+i-1}\}^2\right] \tag{5.42}$$

using $u_{t+i} = y_{t+i} + ku_{t+i-1}$, so that

$$u_{t+i-1} = y_{t+i-1} + k^2y_{t+i-2} + \ldots + k^{i-1}u_t.$$

Note that (5.42) can be written as

$$\frac{1}{(2\pi V/k)^{\ell/2}}\exp\left\{\frac{-k}{2V}\sum_{i=1}^{\ell}(u_{t+i}-u_{t+1-i})^2\right\}.$$

On making the substitution $z_{t+i} = y_{t+i} - (1-k)u_t$, (5.42) becomes a positive quadratic form in z_{t+i}, since

$$y_{t+i} - (1-k)u_{t+i-1} = z_{t+i} - (1-k)(z_{t+i-1} + \ldots + k^{i-2}z_{t+1})$$

so that the joint predictive distribution is multivariate normal with each marginal predictive distribution $f(y_{t+\ell}|y^t)$ univariate normal with mean $E[y_{t+\ell}|y^t] = m_t$. Consequently using any symmetric loss function (5.19) and (5.20) are satisfied. In particular with any step loss

function we must have (5.34) satisfied, which also follows from (5.40), (5.41) giving $k_0 = 1/\{(1-k)^2 + 1\}$.

Indeed this example has all the desirable properties since either from Example 5.4 or using (5.38), (5.39)

$$\theta_{t+2}|y^t \sim N(m_t, V(1-k)(2-k)/k)$$

so that $$p(\theta_{t+2}|y^t) \propto p(\theta_{t+1}|y^t)^{(2-k)^{-1}}$$

and similar results can be obtained for $p(\theta_{t+\ell}|y^t)$.

Example 5.8

If the observations have an exponential distribution with parameter θ_t

$$f(y_t|\theta_t) = \begin{cases} \theta_t \exp(-\theta_t y_t) & y_t > 0 \\ 0 & \text{elsewhere} \end{cases} \qquad (5.43)$$

then the appropriate conjugate posterior for θ_t is the gamma distribution

$$p(\theta_t|y^t) \propto \theta_t^{(1-k)^{-1}} \exp(-\theta_t u_t) \quad \theta_t > 0 \qquad (5.44)$$

where $u_t = \Sigma k^j y_{t-j}$, and $p(\theta_{t+1}|y^t)$ also has a gamma distribution. In this case

$$\exp\{\lambda(u)\} = \Gamma\{(2-k)/(1-k)\}\, u^{-(2-k)(1-k)^{-1}}$$

$$\exp\{\rho(ku)\} = \Gamma\{1/(1-k)\}\, ku^{-(1-k)^{-1}}$$

where $\Gamma(\alpha)$ is the gamma function $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$. Both $\theta_t|y^t$ and $\theta_{t+1}|y^t$ are unimodal with mode $1/\{(1-k)u\}$.

Under this model the predictive distributions are given by

Theorem 5.9

$$f(y_{t+1}|y^t) = (1-k)^{-1} \left(\frac{ku_t}{ku_t + y_{t+1}} \right)^{(1-k)^{-1}} \frac{1}{(ku_t + y_{t+1})} \qquad (5.45)$$

which is a Pareto distribution whilst

$$f(y_{t+2}|y^t) = (k^2u_t)^{(1-k)^{-1}}(1-k)^{-2}\int_{ku_t}^{\infty} z^{-1}(kz+y_{t+2})^{-(2-k)/(1-k)}\,dz. \tag{5.46}$$

Proof

From (5.28) and above

$$f(y_{t+1}|y^t) = \int_0^{\infty} \theta\, e^{-\theta y}\, \theta^{k/(1-k)}\, \frac{(ku)^{1+k/(1-k)}}{\Gamma\{1+k/(1-k)\}}\, e^{-ku\theta}\, d\theta$$

$$= \int_0^{\infty} \theta^{(1-k)^{-1}}\, e^{-\theta(ku+y)}\, \frac{(ku)^{(1-k)^{-1}}}{\Gamma\{1/(1-k)\}}\, d\theta$$

$$= \frac{\Gamma[1+\{1/(1-k)\}]\ (ku)^{(1-k)^{-1}}}{(ku+y)^{\,1+\{1/(1-k)\}}\ \Gamma\{1/(1-k)\}}$$

which since $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ (5.47)

reduces to (5.45). From (5.30)

$$f(y_{t+2}|y^t) = \exp\{\alpha(y_{t+2})\}\,\frac{(ku)^{(1-k)^{-1}}}{\Gamma\{1/(1-k)\}}$$

where from (5.31)

$$\exp\{\alpha(y_{t+2})\} = \int \Gamma(k^*)\ (y_{t+2}+ku)^{-k^*}\,\frac{\Gamma(k^*)\ u^{-k^*}}{\Gamma\{1/(1-k)\}}\ (ku)^{(1-k)^{-1}}dy_{t+1}$$

where $k^*=(2-k)/(1-k)$ and where $u\equiv u_{t+1}=ku_t+y_{t+1}$. If we now denote ku_t+y_{t+1} by $ku+z$ then using (5.47) with the above

$$f(y_{t+2}|y^t) =$$

$$\frac{(ku)^{(1-k)^{-1}}}{(1-k)^2}\int (y+k^2u+kz)^{-k^*}(ku+z)^{-k^*}\,(k^2u+kz)^{(1-k)^{-1}}\,dz$$

$$= \frac{(k^2u)^{(1-k)^{-1}}}{(1-k)^2}\int_0^{\infty} \frac{(y+k^2u+kz)^{-k^*}}{ku+z}\,dz$$

or transforming $= \frac{(k^2u)}{(1-k)^2}(1-k)^{-1} \int_{ku_t}^{\infty} z^{-1}(kz + y_{t+2})^{-k^*} dz.$

We can show that for the predictive distributions

Lemma 5.10 $f(y_{t+\ell}|y^t)$ is a monotonically decreasing function of $y_{t+\ell}$ for all lead times $\ell \geq 1$.

Proof

$$f(y_{t+\ell}|y^t) = \int_0^{\infty} f(y_{t+\ell}|\theta_{t+\ell}) p(\theta_{t+\ell}|y^t) d\theta_{t+\ell}$$

thus

$$\frac{\partial}{\partial y_{t+\ell}} f(y_{t+\ell}|y^t) = \int_0^{\infty} \frac{\partial}{\partial y_{t+\ell}} f(y_{t+\ell}|\theta_{t+\ell}) p(\theta_{t+\ell}|y^t) d\theta_{t+\ell}.$$

But

$$\frac{\partial}{\partial y} f(y|\theta) = -\theta^2 e^{-\theta y} < 0 \quad \text{for all } \theta, y \geq 0 \quad \text{and so}$$

$\frac{\partial f(y_{t+\ell}|y^t)}{\partial y_{t+1}} < 0$, that is $f(y_{t+\ell}|y^t)$ is a strictly decreasing function.

Because of this result, the point forecasts are heavily dependent upon the shape of the loss function used. For example under step loss functions gauge b (defined in (5.15)),

$$y_t(1) = y_t(\ell) = b \tag{5.48}$$

so that (5.19) is satisfied, and (5.21) trivially that $k = 1$. Of course the decisions (5.48) are not particulary sensible since they are independent of the data.

Under such loss functions the expected loss is $1-F(2b)$ where F is the appropriate distribution function. Using (5.45), (5.46) and (5.48) we have in an obvious notation that the expected loss of decision (5.48) at lead time one is

$$E[L\{y_t(1)\}] = \int_{2b}^{\infty} \frac{1}{(1-k)} \left(\frac{ku_t}{(ku_t+y)}\right)^{(1-k)^{-1}} \frac{1}{(y+ku_t)}\, dy$$

$$= \frac{(ku)^{(1-k)^{-1}}}{(ku+2b)^{(1-k)^{-1}}}$$

and for lead time 2

$$E[L\{y_t(2)\}] = \frac{(k^2u_t)^{(1-k)^{-1}}}{(1-k)^2} \int_{2b}^{\infty} dy \int_{ku}^{\infty} \frac{dz}{z(kz+y)^{k^*}}$$

where k^* is as defined above. Using Fubini's theorem to interchange the order of integration simplifies this to

$$E[L\{y_t(2)\}] = \frac{(k^2u_t)^{(1-k)^{-1}}}{(1-k)} \int_{ku}^{\infty} \frac{dz}{z(kz+2b)^{1/(1-k)}} .$$

An interesting point emerges from the following lemma where the loss function is step loss gauge b.

Lemma 5.11

For certain values of k, u_t, b we have

$$E[L\{y_t(2)\}] < E[L\{y_t(1)\}]$$

which says that the uncertainty associated with decisions (5.48) decreases, thus violating Smith's axiom (ii) of §5.3 for a steady evolution (applied to the predictive distributions).

Proof

Provided that $0<k<1$

$$f(y_{t+1}|u_t)(0) = \frac{1}{(1-k)ku}$$

and

$$f(y_{t+2}|u_t)(0) = \frac{(k^2u)^{(1-k)^{-1}}}{(1-k)^2} \int_{ku}^{\infty} \{(kz)^{-(2-k)(1-k)^{-1}}\} z^{-1} dz$$

$$= \frac{(k^2u)^{1/(1-k)}}{(1-k^2)k^{(2-k)/(1-k)}} \left[\frac{1-k}{(2-k)\ z^{(2-k)/(1-k)}} \right]_{\infty}^{ku}$$

$$= \frac{1}{(2-k)(1-k)k^2u} \ .$$

But $\frac{1}{(2-k)(1-k)k^2u} > \frac{1}{(1-k)ku}$ since $1 > 2k-k^2$

thus $f(y_{t+2}|y^t)(0) > f(y_{t+1}|y^t)(0)$.

Since both of these functions are continuous, there is some b such that on $[0,2b)$, $f(y_{t+2}|y^t) > f(y_{t+1}|y^t)$ so that

$$F_{y_t(2)}(2b) > F_{y_t(1)}(2b)$$

which implies that

$$1 - F_{y_t(2)}(2b) < 1 - F_{y_t(1)}(2b)$$

or

$$E[\,L\{y_t(2)\}\,] < E[\,L\{y_t(1)\}\,] \ .$$

If we consider the usual quadratic loss functions then the predictors are just the conditional mean. The mean of the Pareto distribution (5.45) is $(1-k)u_t$, so that

$$y_t(1) = m_t = (1-k)u_t$$

which is the traditional EWMA.

For the two-step ahead predictor

$$E[y_{t+2}|y^t] = E\{E[y_{t+2}|y^{t+1}]\}$$

$$= E[(1-k)u_{t+1}]$$

$$= E[(1-k)(y_{t+1} + ku_t)]$$

$$= (1-k)\{(1-k)u_t + ku_t\}$$

$$= (1-k)u_t$$

so that by an inductive argument

$$E[y_{t+\ell}|y^t] = m_t$$

or $$y_t(\ell) = y_t(1) = m_t.$$

The associated expected loss is the conditional variance

$$\int (y_{t+\ell} - m_t)^2 f(y_{t+\ell}|y^2)\, dy_{t+\ell}$$

so that for the uncertainty to increase we require

$$E[y_{t+2}{}^2|y_t] > E[y_{t+1}{}^2|y_t] \qquad (5.49)$$

We shall now prove this :

Theorem 5.12

For this model under quadratic loss the uncertainty is increasing between lead times 1 and 2, that is

$$E[L\{y_t(2)\}] > E[L\{y_t(1)\}] .$$

Proof

$E[y_{t+1}{}^2|y^t]$ is given in most text-books, however directly we find

$$E[y_{t+1}{}^2|y^t]$$

$$= \frac{1}{(1-k)} \int_0^\infty y^2 \left(\frac{ku_t}{ku_t+y}\right)^{1/(1-k)} \frac{1}{(ku+y)}\, dy$$

$$= \frac{1}{(1-k)} \int_0^\infty \{ku_t+y)^2 - (ku_t)^2 - 2ku_t y\}\left(\frac{ku_t}{ku_t+y}\right)^{1/(1-k)} \frac{1}{(ku_t+y)}\, dy$$

$$= \frac{(ku)^{1/(1-k)}}{(1-k)} \left[\frac{1-k}{(2k-1)(ku)^{(2k-1)/(1-k)}}\right] - (ku)^2 - 2ku(1-k)u$$

provided that $2k - 1 > 0$, ie $k > \frac{1}{2}$,

$$= \frac{(ku)^2}{2k-1} - (ku)^2 - 2ku^2(1-k) = \frac{2ku^2}{2k-1}(1-k)^2$$

$$= \frac{2km^2}{2k-1} . \qquad (5.50)$$

Now $E[y_{t+2}^2 | u_t] = \frac{(k^2u)^{1/1-k}}{(1-k)^2} \int_0^\infty \int_{ku}^\infty \frac{y^2 \, dz \, dy}{(kz+y)^{(2-k)/(1-k)}} z$

which on changing the order of integration and using (5.50)

$$= \frac{(k^2u)^{1/(1-k)}}{(1-k)^2} \int_{ku}^\infty \frac{dz}{z} \left\{ \frac{(1-k)}{(kz)^{1/(1-k)}} \frac{2kz^2(1-k)^2}{2k-1} \right\}$$

$$= 2\frac{(ku)^{1/(1-k)}}{2k-1} k(1-k) \int_{ku}^\infty \frac{dz}{z^{k/(1-k)}}$$

$$= 2(ku)^{1/(1-k)} \frac{k(1-k)}{2k-1} \left\{ \frac{(1-k)}{(2k-1)(ku)^{(2k-1)/(1-k)}} \right\}$$

provided that $k > \frac{1}{2}$,

$$= \frac{2(1-k)^2(ku)^2k}{(2k-1)^2} = \frac{2k^3m^2}{(2k-1)^2} .$$

But $\frac{2k^3m^2}{(2k-1)^2} > \frac{2km^2}{2k-1}$ because $k^2 > 2k - 1$, so that (5.49) is satisfied, and thus $E[L\{y_t(2)\}] > E[L\{y_t(1)\}]$.

Consequently under quadratic loss this model appears to emulate most of the features of the Harrison-Stevens steady model.

This example, like the last, illustrates the important point that the steady evolution of the system parameter is essentially a one-step phenomenon. We have from (5.39) that

$$\exp\{\beta(\theta_{t+2})\} = \int_0^\infty \frac{e^{-\theta y k}\,\Gamma\{(2-k)/(1-k)\}\,(ku_t+y)^{-(2-k)/(1-k)}}{\Gamma\{1/(1-k)\}\,\{k(ku_t+y)\}^{-1/(1-k)}}\,dy$$

$$= \int_0^\infty \frac{k^{1/(1-k)}}{(1-k)}\; e^{-\theta k y}\,(ku_t + y)^{-1}\,dy$$

so that from (5.38)

$$p(\theta_{t+2}|y^t) = \frac{e^{-k^2u\theta}\;\theta^{k/(1-k)}\;(k^2u_t)^{1/1-k}}{(1-k)\;\Gamma\{1/(1-k)\}} \int_0^\infty e^{-\theta y k}(ku_t+y)^{-1}\,dy.$$

On transforming the integral using $z=\theta k(y+ku_t)$ we have

$$p(\theta_{t+2}|y^t) = \frac{\theta^{k/(1-k)}\;(k^2u_t)^{1/(1-k)}}{(1-k)\;\Gamma\{1/(1-k)\}}\; E_1(k^2u_t\theta) \qquad (5.51)$$

where $E_1(x) = \int_x^\infty \frac{e^{-z}}{z}\,dz$ is the exponential integral whose properties are given in Abramowitz and Stegun (1965, p228).

As remarked upon earlier, the conditional densities $p(\theta_{t+1}|y^t)$ and $p(\theta_t|y^t)$ are unimodal with mode at $\{u_t(1-k)\}^{-1}$ provided that $1>k>0$. Now

$$\frac{\partial}{\partial\theta_{t+2}}\,p(\theta_{t+2}|y^t) = \frac{(k^2u_t)^{1/(1-k)}\;\theta^{(2k-1)/(1-k)}}{(1-k)\;\Gamma\{1/(1-k)\}}\;x \qquad (5.52)$$

$$x\;\left\{\frac{k}{1-k}\;E_1(k^2u_t\theta) - e^{-k^2u_t\theta}\right\};$$

from Abramowitz and Stegun(ibid),

$$\frac{1}{x+1} < e^x\,E_1(x) \leq \frac{1}{x}$$

so that if $\theta > \{(1-k)ku_t\}^{-1}$ then

$$e^{k^2u_t\theta}\;E_1(k^2u_t\theta) < \frac{1-k}{k}$$

and so from (5.52) $p(\theta_{t+2}|y^t)$ is decreasing. On the other hand, if $\theta < (2k-1)/\{(1-k)k^2u_t\}$ with $k>\frac{1}{2}$ then

$$e^{k^2u\theta}\, E_1(k^2u\theta) > \frac{1}{1 + \frac{2k-1}{1-k}} = \frac{1-k}{k}$$

and $p(\theta_{t+2}|y^t)$ is increasing. The density $p(\theta_{t+2}|y^t)$ given by (5.51) is continuous, and so from the last two results has at least one mode $\hat{\theta}_t$ satisfying

$$\frac{2k-1}{(1-k)k^2u} < \hat{\theta}_t < \frac{1}{(1-k)ku} .$$

In general we will not have $\hat{\theta}_t = \{(1-k)u_t\}^{-1}$ since for almost all k

$$\frac{k}{1-k}\, E_1\left(\frac{k^2}{1-k}\right) \neq e^{k^2/(1-k)}$$

so that the mode of $p(\theta_{t+2}|y^t)$ will not be the same as that for $p(\theta_{t+1}|y^t)$, a fact which precludes any relationship of the form

$$p(\theta_{t+2}|y^t) \propto T\{p(\theta_{t+1}|y^t)\}$$

for a convex function T.

For this Gamma-exponential model it can be seen (from Example (5.6) or directly) that

$$\frac{\phi^{\frac{k}{1-k}}\, e^{-ku\phi}\, (ku)^{-(1-k)^{-1}}}{\Gamma\{1/(1-k)\}} = \int \frac{p(\phi|\theta)\, \theta^{(1-k)^{-1}}\, e^{-u\theta}\, u^{\frac{2-k}{1-k}}}{\Gamma(\frac{2-k}{1-k})}\, d\theta$$

is satisfied by

$$p(\phi|\theta) = \left(\frac{k}{1-k}\right)\left(\frac{1}{\theta}\right)\left(\frac{k\phi}{\theta}\right)^{\frac{k}{1-k}} \qquad 0 < k\phi < \theta \qquad (5.53)$$

which is a special case of Bather's (1965) example. $p(\phi|\theta)$ satisfies requirements (i) and (ii) introduced after (5.32), and indeed $\frac{k\phi}{\theta}$ has a Beta distribution. Therefore in this

example $p(\theta_t|y^t)$ and $p(\theta_{t+1}|y^t)$ form invariant conditional densities and so all of Bather's theory carries across.

To summarise, this example is interesting because (1) it can be described by an observation equation on a Markov chain whose transition density is given by (5.53); (2) under quadratic loss it is a steady model in that (5.19), (5.20) hold with the forecasts given by the EWMA (5.14), moreover the uncertainty associated with such decisions is increasing with each lead time as required.

Example 5.13 The Gamma-Gamma Model

Example 5.8 can be simply extended to the case where y has a gamma distribution

$$f(y|\theta) = \frac{1}{\Gamma(\alpha)}\theta^{\alpha}y^{\alpha-1}e^{-\theta y}$$

with $a(\theta) = -\theta$, $b(y)=y$, $c(\theta)=\alpha\log\theta$, $d(y)=(\alpha-1)\log y - \log\Gamma(\alpha)$. Under conjugate analysis

$$\theta_t|y^t \propto e^{-\theta_t\gamma_t}\theta^{\alpha\delta_t}$$

so that again the system parametres have gamma distributions, with limiting values

$$p(\theta_t|y^t) \propto \begin{cases} e^{-u_t\theta_t}\theta^{\{\alpha/(1-k)\}} & \theta_t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

For this example

$$\exp\{\lambda(u) = \Gamma\left(1+\frac{\alpha}{1-k}\right)u^{-\left(\frac{\alpha}{1-k}+1\right)}$$

$$\exp\{\rho(ku)\} = \Gamma\left(1+\frac{k\alpha}{1-k}\right)(ku)^{-\left(1+\frac{k\alpha}{1-k}\right)}$$

with predictive distribution from (5.28)

$$f(y_{t+1}|y^t) = \frac{y^{\alpha-1}\ \Gamma(1+\frac{\alpha}{1-k})(ku_t+y)^{-(\frac{\alpha}{1-k}+1)}}{\Gamma(\alpha)\ \Gamma(1+\frac{k\alpha}{1-k})(ku_t)^{-(1+\frac{k\alpha}{1-k})}}$$

$$= \frac{y^{\alpha-1}\ (ku_t)^{\frac{\alpha k}{1-k}+1}}{(y+ku_t)^{\frac{\alpha}{1-k}+1}\ Be(\alpha,\frac{\alpha k}{1-k}+1)} \qquad (5.54)$$

where $Be(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, so that (5.54) is an Inverse-Beta distribution with mean $(1-k)u_t$ and mode $\frac{(1-k)ku_t(\alpha-1)}{2+k(\alpha-2)}$ and where, as in the previous example, $u_t = \sum k^j y_{t-j}$. The predictive distributions soon become complicated, however under quadratic loss

$$y_t(\ell) = E[y_{t+\ell}] = m_t = (1-k)u_t.$$

As in Example 5.8 the steady evolution of the system parameter can be expressed by means of a Beta transition density from θ_t to θ_{t+1}.

Examples 5.8 and 5.13 illustrate a general point: suppose that we choose to work within the framework of quadratic loss functions, then for a steady model we require at least

$$E[y_{t+2}|y^t] = E[y_{t+1}|y^t]$$

or $E_{y_{t+1}}\{E[y_{t+2}|y^{t+1}]\} = E[y_{t+1}|y^t]$.

We mentioned earlier the desirability of a stochastic form of stability, which in this context implies that there is some function f such that

$$E[y_{t+1}|y^t] = f(u_t).$$

Our requirement for a steady model then becomes

$$\underset{y_{t+1}}{E}\,[\, f(u_{t+1})\,] = f(u_t).$$

In the case of linear connecting functions, $u_{t+1} = ku_t + b(y_{t+1})$ so that we then require

$$\underset{y_{t+1}}{E}\,[\, f(ku_t + b(y_{t+1}))] = f(u_t).$$

Under the 'natural parametrization' this becomes

$$\underset{y_{t+1}}{E}\,[\, f(ku_t + y_{t+1})\,] = f(u_t). \qquad (5.55)$$

The simplest examples of such functions are when f is a linear function, $f(u_t) = au_t + b$ say, then (5.55) holds if

$$E[\, a(ku_t + y_{t+1}) + b\,] = au_t + b$$

which is true if

$$aku_t + a^2u_t + ab = au_t$$

or

$$(k+a)u_t + b = u_t.$$

For this to be true for all u_t we require $b = 0$, $a = (1-k)$ which gives

$$E[y_{t+1}|y^t] = (1-k)u_t. \qquad (5.56)$$

This relation holds for Examples 5.8, 5.13 and the normal model of Example 5.7.

Example 5.14 The Beta-Binomial Model

We now give an example where neither (5.19) nor (5.20) is satisfied. It is assumed that the observation variable has a binomial distribution

$$f(y_t|\theta_t) = \begin{cases} \binom{n}{y_t} \theta_t^{y_t}(1-\theta_t)^{n-y_t} & y_t = 0,1 \ldots n \\ 0 & \text{otherwise} \end{cases} \quad (5.57)$$

with n known. Again $b(y_t) = y_t$, so that $u_t = \sum k^j y_{t-j}$ giving under conjugate analysis in the steady state.

$$p(\theta_t|y^t) \propto \theta_t^{u_t}(1-\theta_t)^{n(1-k)^{-1}-u_t} \qquad 0 < \theta_t < 1$$
$$= 0 \qquad \text{otherwise}$$

which is a Beta distribution. The other requisite quantities are $d(y) = \log\binom{n}{y}$

$$\exp\{\lambda(u)\} = B\{u+1, n(1-k)^{-1}-u+1\} \quad (5.58)$$

$$\exp\{\rho(ku)\} = B\{ku+1, nk(1-k)^{-1}-ku+1\}. \quad (5.59)$$

The predictive distribution is from (5.28)

$$f(y_{t+1}|y^t) = \binom{n}{y_{t+1}} \exp\{\lambda(ku+y_{t+1}) - \rho(ku_t)\} \quad y_{t+1} = 0,1 \ldots n$$

which on substituting (5.58) and (5.59) becomes a Beta binomial distribution, and from (5.30)

$$f(y_{t+2}|y^t) = \binom{n}{y_{t+2}} \exp\{\alpha(y_{t+2}) - \rho(ku_t)\} \qquad y_{t+1} = 0 \ldots n$$

where

$$\exp\{\alpha(y_{t+2})\} = \sum_{y_{t+1}=0}^{n} \binom{n}{y_{t+1}} \exp\{\lambda(u_{t+2}) + \lambda(u_{t+1}) - \rho(ku_{t+1})\}$$

with $u_{t+2} = ku_{t+1} + y_{t+2}$, $u_{t+1} = ku_t + y_{t+1}$.

The predictive distributions are discrete in this example, so that step loss functions are no longer so appropriate, and the question arises as to what is a suitable loss function. Under quadratic loss, the conditional mean is

$$E[y_{t+1}|u_t] = E_{\theta_{t+1}|u_t}[n\theta_{t+1}]$$

$$= \frac{n(1-k)(ku_t+1)}{kn(1-k)^{-1}+2}$$

$$= \phi(ku_t+1)$$

where the constant $\phi = \frac{n(1-k)}{nk+2-2k}$.

It follows from (5.55) and (5.56) that in general $E[y_{t+2}|u_t] \neq E[y_{t+1}|u_t]$; in fact directly

$$E[y_{t+2}|u_t] = \phi E[k(y_{t+1}+ku_t)+1]$$

$$= \phi\{k^2u_t+1+k\phi(ku_t+1)\}$$

$$= \phi\{(\phi+1)(k^2u_t) + k\phi+1\}$$

so that $E[y_{t+2}|u_t] \neq E[y_{t+1}|u_t]$ unless

$$(\phi+1)k^2u_t+k\phi = ku_t.$$

This is true if $(n+2-2k)ku_t + n(1-k) = (nk+2-2k)u_t$

which implies $(2-2k)u_t(k-1) + n(1-k) = 0$

giving $u_t = \frac{n}{2(1-k)}$.

It is not easy to apply standard loss functions to such predictive distributions. One estimate we might use is the mode of $y_{t+1}|u^t$, which is the integer part of $\binom{n+1}{n}(1-k)u_t$.

5.5 Other Models and Extensions

We can of course apply the steady evolution to obtain tractable results with any single parameter distribution of

of exponential form and its conjugate. Although all the examples give have satisfied $b(y)\equiv y$ there is no need to apply this restriction; for example we could have the observations distributed as a Rayleigh distribution in which case $b(y)=y^2$. The examples have been chosen to illustrate some of the points made earlier.

We can think of all these examples as being a particular subset of the univariate class of 'conjugacy state-space models'. These can be represented as

$$f(y_t|\theta_t) = \exp\{a(\theta_t)b(y_t) + c(\theta_t) + d(y_t)\} \qquad (5.60)$$

$$p(\theta_t|y^t) \propto \exp\{\gamma_t a(\theta_t) + \delta_t c(\theta_t)\} \qquad (5.61)$$

$$p(\theta_{t+1}|y^t) \propto \exp\{\alpha_t a(\theta_t) + \beta_t c(\theta_t)\} \qquad (5.62)$$

together with a mapping $\tau : (\gamma_t, \delta_t) \to (\alpha_t, \beta_t)$ (5.63)

so that Smith's models correspond to τ being the mapping

$$(\gamma_t, \delta_t) \to (k\gamma_t, k\delta_t).$$

Then applying Bayes' theorem

$$p(\theta_{t+1}|y^{t+1}) \propto f(y_{t+1}|\theta_{t+1})p(\theta_{t+1}|y^t)$$

and so from (5.60), (5,62)

$$\propto \exp\{\gamma_{t+1}a(\theta_{t+1}) + \delta_{t+1}c(\theta_{t+1})\}$$

where $$\gamma_{t+1} = \alpha_t + b(y_{t+1}) \qquad (5.64)$$

$$\delta_{t+1} = \beta_t + 1 \qquad (5.65)$$

Under such a system we can look at some kind of steady evolution (suitably defined), or other types of behaviour.

In the above, y_t can be a vector, and the extension to vector parameters is obtained by considering

$$f(\underline{y}_t|\underline{\theta}_t) = \exp\{\Sigma a_i(\underline{\theta}_t)b_i(\underline{y}_t) + \underline{c}(\underline{\theta}_t) + \underline{d}(\underline{y}_t)\} \qquad (5.66)$$

in place of (5.60), which with the appropriate conjugate priors gives the set of multiparameter state-space models. Indeed, although this chapter has concentrated on the unvariate case, the theory applies equally well to multivariate multiparameter models, provided that the obvious extensions are made; for example all the theory of Section 5.3 is directly applicable to multivariate observations.

Smith (1981) considers two types of multiparameter models. The first is the 'Symmetric Multivariate Power Steady Model', SMPSM where $\underline{\theta}$ evolves through (5.17) with $\underline{\theta}$ a vector, and the second is the 'Stacked Steady Model' with $(\theta_i|\theta_{i+1} \ldots \theta_n)$ evolving as a SMPSM with parameter k_i, $0<k_i<k_{i+1}\leq 1$. The predictive consequences of both of these models can be analysed using the techniques of Section 5.3 by suitably generalising the equations, for instance by using the generalised exponential family (5.66) with conjugate priors, thus giving a particular class of conjugate state-space models.

Similar remarks apply to the predictive distributions of multiparameter models as were made for the single parameter models. For example the forecasts can be heavily dependent upon the loss function used, and for univariate observations the relations (5.19) - (5.20) need not hold. These are consequences of the fact that the steady evolution is for the parameters rather than the observations.

CHAPTER 6

GENERAL EQUIVALENCE THEOREMS FOR DLMs

6.1 Discussion

It is well known that certain equivalences exist between different types of models used to describe Time Series, or models used in forecasting. Before proceeding, it is worth specifying what we mean by equivalence:

Definition 6.1

We say that two forecasting systems are predictor equivalent if their forecast functions coincide for all time, that is $y_t(k)$ is the same for all t and lead times k given data up to time t.

A slightly stronger definition is that two forecasting systems are (second-order) completely predictor equivalent if they are predictor equivalent and if in addition the uncertainty associated with each forecast as expressed by the variance is the same for the two systems, again for all t and k.

This is not such a useful definition since within a Bayesian framework it only really makes sense with quadratic loss functions. The strongest relationship between models that describe Time Series is

Definition 6.2

Two descriptions of a time series $\{x_t\}$ are said to be model equivalent if they induce the same joint probability distribution over $\{x_t\}$ for all t.

Although this appears to be the most important type of equivalence it is inadequate in a forecasting situation

because firstly we do not need to postulate a probability model to define a legitimate forecasting system, and secondly because model equivalence does not imply predictor equivalence. For example in a Bayesian context we would need to specify equivalent loss functions. Another example is provided by the linear model in statistics, where the x_t depend upon a parameter θ. Then the predictor of x_{t+1} given data up to time t will depend upon the way we have estimated θ.

Several authors - for example Priestley (1980) - have commented on the fact that ARMA models can be expressed in a state-space form (DLM), showing model equivalence; in effect this is discussed in Section 3.6 where (3.13) represents an ARMA model if the u_i are regarded as error terms. There are many equivalent state-space representations and a 'canonical' description is provided by Theorem 3.31, which has similarities with the canonical description we develop later. Akaike (1974) only considers stationary ARMA models and uses the Wold decomposition theorem to obtain a state-space representation, where the system vector comprises conditional expectations at various lead times of the observations given data up to time t and where the error terms depend upon the innovation of y_t, that is the difference between y_t and the 1-step ahead predictor of y_t at time t-1. Because of the particular way the state-space is defined, this embodies a type of predictor equivalence - provided that conditional expectations are used as forecasts, but no mention is made of the Kalman filter.

We shall be concerned with predictor equivalence. In

this context McKenzie (1976) showed that certain traditional forecasting schemes based on EWMA's, such as Holt-Winter's method, are predictor equivalent to particular Box-Jenkins' ARIMA models. Godolphin and Harrison (1975) looked at methods having a polynomial forecast function and showed that they were all equivalent to IMA models. In particular the Markov polynomial models (Harrison, 1967) which are a type of DLM fall in this category, but are limited by being a strict subset of the IMA models. Godolphin and Stone (1980) considered polynomial projecting models based upon DLMs and the Kalman filter, deriving conditions for the DLMs not to be limited. Stone (1982) extended this work, looking at equivalence for more general forecast functions.

In this chapter Section 2 stresses the role of observability, while Section 3 examines the equivalence between DLMs in the steady state and ARIMA models, giving general results anticipated in the work of Godolphin and Stone. Section 4 shows that time-varying DLMs are equivalent to time-dependent ARIMA models, and Section 5 relates model equivalence to predictor equivalence. Section 6 examines the structural properties of DLM.

6.2 Non-Derogatory Models and DLMs

Consider the general univariate time-invariant DLM

$$y_t = \underline{F}\,\underline{\theta}_t + v_t \qquad (6.1)$$

$$\underline{\theta}_t = \underline{G}\,\underline{\theta}_{t-1} + \underline{w}_t \qquad (6.2)$$

where y_t is a scalar and where from now on we use assumptions (4.5), (4.6) unless stated otherwise. Provided that we use a symmetric loss function the forecast function at lead time ℓ is given by (4.25)

$$y_t = \underline{F}\,\underline{G}^{\ell}\,m_t \qquad (6.3)$$

where $\underline{m}_t$ is obtained from (4.9) - (4.11).

Since $\underline{G}$ is a square matrix and $\underline{F} \in \mathbb{R}^n$, there is a positive integer k such that $\underline{F}\underline{G}$, ... $\underline{F}\underline{G}^k$ are linearly independent row vectors, whilst $\underline{F}\underline{G}$, ... $\underline{F}\underline{G}^{k+1}$ are linearly dependent. Let us call this k the horizon of $\{\underline{F},\underline{G}\}$. It then follows that for any ℓ $\underline{F}\underline{G}^{\ell}$ can be expressed as a linear combination of $\{\underline{F}\underline{G}, \ldots \underline{F}\underline{G}^k\}$ so that knowledge of k, $\underline{m}_t$ and $\{\underline{F}\underline{G}, \ldots \underline{F}\underline{G}^k\}$ completely determines the forecast function.

This is closely related to a more familiar concept in matrix theory:

Definition 6.3

For any row vector $\underline{F} \in \mathbb{R}^n$ and square nxn real matrix $\underline{G}$, the $\underline{G}$-order of $\underline{F}$ is the monic polynomial of least degree, $h_F(x)$ which satisfies

$$\underline{F}h_F(\underline{G}) = 0$$

where $h_F(x)$ is a member of the polynomial ring over the reals. Therefore $h_F(x)$ is the minimal polynomial of the linear transformation $\underline{G}_F$ induced by the $\underline{G}$-cyclic subspace generated by $\underline{F}$ (under the fundamental isomophism between matrices and linear transformations). See Birkhoff and Maclane (1965, chapter 10, §7). The following properties follow

Lemma 6.4

(i) If g(x) is the minimal polynomial of $\underline{G}$ then $h_F(x) | g(x)$ for all $F \in \mathbb{R}^n$, where our notation means that $h_F(x)$ divides g(x).

(ii) There is some $\underline{F}$ such that $h_F(x) \equiv g(x)$.

(iii) $h_F(x) \equiv g(x) \equiv h_Y(x)$ if and only if $\underline{Y} = \underline{F}\Gamma(G)$ where $\Gamma(x)$ is coprime with $g(x)$, provided that $g(x)$ is a polynomial of the same degree as $\underline{G}$.

Proof

Parts (i) and (ii) are derived in Birkhoff and Maclane (ibid). It is also possible to adapt the argument of Theorem 19 of Birkhoff and Maclane (p 303) to derive part (iii) of the present theorem, however the following direct proof appears to lend some insight to the result. Suppose $\underline{Y} = \underline{F}\Gamma(G)$, where $\Gamma(x)$ is coprime with $g(x)$, then

$$\underline{F}\Gamma(\underline{G})h_Y(\underline{G}) = 0$$

which implies $h_F | \Gamma(\underline{G})h_Y(\underline{G})$

so that $h_F | h_Y(\underline{G})$

by assumption of $\Gamma(x)$ comprime with $g(x)$.

But $\underline{Y}h_F(\underline{G}) = 0$ implies that $h_Y | h_F$, so $h_F = h_Y$.

Conversely since $g(x)$ has order n then the linearly independent vectors $\underline{F}, \underline{F}\underline{G} \ldots \underline{F}\underline{G}^{n-1}$ are a basis for $\mathbb{R}^n$, so that

$$\underline{Y} = \underline{F}\Gamma(\underline{G})$$

for some polynomial $\Gamma(x)$. Suppose $g(x) = \Gamma(x)f(x)$, then

$$\underline{Y}f(\underline{G}) = \underline{F}\Gamma(\underline{G})f(\underline{G}) = \underline{F}g(\underline{G}) = 0$$

so $h_Y(x) = g | f$ which is a contradiction, therefore g, Γ are coprime.

The decomposition Theorem 3.26 states that (6.1), (6.2) is algebraically equivalent to

$$y_t = (0 \quad \underline{F}_o)\begin{pmatrix}\underline{\theta}_u \\ \underline{\theta}_o\end{pmatrix} + v_t \qquad (6.4)$$

$$\begin{pmatrix} \underline{\theta}_u \\ \underline{\theta}_o \end{pmatrix} = \begin{pmatrix} \underline{G}_{uu} & \underline{G}_{uo} \\ 0 & \underline{G}_{oo} \end{pmatrix} \begin{pmatrix} \underline{\theta}_u \\ \underline{\theta}_o \end{pmatrix} + \begin{pmatrix} \underline{w}_{ut} \\ \underline{w}_{ot} \end{pmatrix} \qquad (6.5)$$

where $\{\underline{F}_o, \underline{G}_{oo}\}$ is observable. That is we have decomposed the system into its unobservable and observable parts. Section 4.5 assures us that (6.1), (6.2) is predictor equivalent to the above system; however we can derive a stronger result:

Theorem 6.5

System (6.4), (6.5) and hence (6.1), (6.2) is predictor equivalent both in terms of y_t and $\underline{\theta}_o$ to the observable subsystem

$$y_t = \underline{F}_o \underline{\theta}_o + v_t \qquad (6.6)$$

$$\underline{\theta}_o = \underline{G}_{oo} \underline{\theta}_o + \underline{w}_t \qquad (6.7)$$

Proof

Using the Kalman updating equations (4.9) - (4.11) and the relation

$$\underline{m}_{ot} = (\underline{0}_{oxu} , \underline{I}_{uxu}) \begin{pmatrix} \underline{m}_{ut} \\ \underline{m}_{ot} \end{pmatrix} = \underline{R} \begin{pmatrix} \underline{m}_{ut} \\ \underline{m}_{ot} \end{pmatrix} \quad \text{say}$$

where $\underline{0}_{oxu}$ is the oxu zero matrix and $\underline{I}_{uxu}$ the u x u identity, then

$$\underline{m}_{ot} = \underline{R} \left[\begin{pmatrix} \underline{G}_{uu} & \underline{G}_{uo} \\ \underline{0} & \underline{G}_{oo} \end{pmatrix} \begin{pmatrix} \underline{m}_{ut-1} \\ \underline{m}_{ot-1} \end{pmatrix} + \underline{A}_t \left\{ y_t - (0 \quad \underline{F}_o) \begin{pmatrix} \underline{G}_{uu} & \underline{G}_{uo} \\ \underline{0} & \underline{G}_{uu} \end{pmatrix} \begin{pmatrix} \underline{m}_{ut-1} \\ \underline{m}_{ot-1} \end{pmatrix} \right\} \right]$$

$$= \underline{G}_{oo} \underline{m}_{ot-1} + \underline{R} \underline{A}_t (y_t - \underline{F}_o \underline{G}_{oo} \underline{m}_{ot-1}).$$

But

$$\underline{R}\underline{A}_t = \underline{R} \begin{pmatrix} \underline{P}_{uu} & \underline{P}_{uo} \\ \underline{P}_{uo} & \underline{P}_{oo} \end{pmatrix} \begin{pmatrix} 0 \\ \underline{F}_o^T \end{pmatrix} \left\{ (0 \quad \underline{F}_o) \begin{pmatrix} \underline{P}_{uu} & \underline{P}_{uo} \\ \underline{P}_{uo}^T & \underline{P}_{oo} \end{pmatrix} \begin{pmatrix} 0 \\ \underline{F}_o^T \end{pmatrix} + V \right\}^{-1}$$

$$= \underline{P}_{oo}\,\underline{F}_o^T\{\underline{F}_o\,\underline{P}_{oo}\,\underline{F}_o^T + V\}^{-1} \qquad (6.8)$$

where

$$\underline{P}_{oo} \text{ is } \underline{R}\begin{pmatrix}\underline{P}_{uu} & \underline{P}_{uo}\\ \underline{P}_{ou} & \underline{P}_{oo}\end{pmatrix}\underline{R}^T = \underline{R}\left[\begin{pmatrix}\underline{G}_{uu} & \underline{G}_{uo}\\ 0 & \underline{G}_{oo}\end{pmatrix}\begin{pmatrix}\underline{C}_{uu} & C_{uo}\\ \underline{C}_{uo}^T & \underline{C}_{oo}\end{pmatrix}\begin{pmatrix}\underline{G}_{uu}^T & 0\\ \underline{G}_{uo}^T & G_{oo}^T\end{pmatrix} + \begin{pmatrix}\underline{W}_{uu} & \underline{W}_{uo}\\ \underline{W}_{ou} & \underline{W}_{oo}\end{pmatrix}\right]\underline{R}^T$$

$$\text{so } \underline{P}_{oo} = \underline{G}_{oo}\,\underline{C}_{oo}\,\underline{G}_{oo}^T + \underline{W}_{oo} \qquad (6.9)$$

$$\text{thus } \underline{m}_{ot} = \underline{G}_{oo}\,\underline{m}_{ot-1} + \underline{P}_{oo}\,\underline{F}_o^T\{\underline{F}_o\,\underline{P}_{oo}\underline{F}_o^T + V\}^{-1}(y_t - \underline{F}_o\underline{G}_{oo}\underline{m}_{ot-1})$$

$$\text{Now } C_t = \begin{pmatrix}\underline{C}_{uu} & \underline{C}_{uo}\\ \underline{C}_{uo}^T & \underline{C}_{oo}\end{pmatrix} = (\underline{I} - \underline{A}_t\underline{F})\underline{P}_t \qquad (6.10)$$

$$\text{thus } \underline{C}_{oo_t} = \underline{R}\,\underline{C}_t\,\underline{R}^T = \underline{P}_{oo} - \underline{P}_{oo}\,\underline{F}_o^T\{\underline{F}_o\,\underline{P}_{oo}\,\underline{F}_o^T + V\}^{-1}\underline{F}_o\,\underline{P}_{oo}\,. \qquad (6.11)$$

But (6.8) - (6.11) are exactly the recursions obtained from the subsystem (6.6), (6.7). Moreover, with a slight abuse of notation

$$\underline{m}_{o,k} = E[\,\underline{\theta}_{ot+k}|y^t\,] = \underline{R}\,E\left[\begin{pmatrix}\underline{\theta}_{u\,t+k}\\ \underline{\theta}_{o\,t+k}\end{pmatrix}\middle| y^t\right] = \underline{R}\,\underline{G}E\left[\begin{pmatrix}\theta_{u\,t+k-1}\\ \underline{\theta}_{o\,t+k-1}\end{pmatrix}\middle| y^t\right] = \underline{G}_{oo}\,\underline{m}_{o,\,t+k-1} \qquad (6.12)$$

$$\text{and } \underline{C}_{o,k} = \mathrm{Var}(\underline{\theta}_{o,t+k}|y^t) = \underline{R}\left[\mathrm{Var}\begin{pmatrix}\underline{\theta}_{u,t+k}\\ \underline{\theta}_{o,t+k}\end{pmatrix}\middle| y^t\right]\underline{R}^T = \underline{R}\left[\underline{G}\ \mathrm{Var}\begin{pmatrix}\theta_{u,t+k-1}\\ \theta_{o,t+k-1}\end{pmatrix}\middle| y^t\right)\underline{G}^T + \underline{W}\Bigg]\underline{R}$$

$$= \underline{G}_{oo}\, \underline{C}_{oo\ k-1}\, \underline{G}_{oo}^{T} + \underline{W}_{oo}. \qquad (6.13)$$

Moreover, we can similarly show

$$E[y_{t+k}|y^t] = \underline{F}_o\, \underline{m}_{o,k} \qquad (6.14)$$

$$Var(y_{t+k}|y^t) = \underline{F}_o\, \underline{C}_{oo,k}\, \underline{F}_o^{T} + \underline{V} \qquad (6.15)$$

so that since (6.12) - (6.15) are the predictive recursions for the subsystem we have completed the proof.

In fact we have used the decomposition theorem which applies to both vector and time varying case. Consequently we have the completely general

Corollary 6.6

Dynamic linear models are predictor equivalent to their observable subsystems.

This motivates us to make the following assumption concerning $\underline{G}$, namely that the only set of scalars $(C_o, \ldots C_{n-1})$ satisfying

$$C_o \underline{I}_n + \sum_{i=1}^{n-1} C_i\, \underline{G}^i = \underline{0}_{nxn} \qquad (6.16)$$

is the set $(C_o, \ldots C_{n-1}) = \underline{0}_n^{T}$, where $\underline{0}_{nxn}$ is an nxn zero matrix and $\underline{0}_n$ the zero n-vector. This is equivalent to saying that the minimal polynomial has degree n and so is equal to the characteristic polynomial, (up to units in the ring of polynomials over the real field), a statement which follows from (6.6) and the Cayley-Hamilton theorem. This last fact is the definition of a non-derogatory matrix. The importance of this is seen from Lemma 6.4 (ii), Theorem 6.5 and the following

Theorem 6.7

If $\{\underline{F},\underline{G}\}$ is observable, that is $\{\underline{F},\underline{FG} \ldots \underline{FG}^{n-1}\}$ are

linearly independent with $\underline{G}$ an n x n matrix, then $\underline{G}$ is non-derogatory.

The proof follows from Lemma 6.4 (i) and the definition of a minimal polynomial.

We draw the following implications

Lemma 6.8

(i) $\underline{G}$ is similar to the companion matrix $\underline{G}_n^*$ given by

$$\underline{G}_n^* = \begin{pmatrix} \underline{0}_{1xn-1} & \underline{I}_{n-1} \\ -\phi_n & -\underline{\phi}_{n-1}^T \end{pmatrix} \tag{6.17}$$

where $\underline{\phi}_{n-1}^T = (\phi_{n-1}, \ldots \phi_1)$ and $\phi_1, \ldots \phi_n$ are the coefficients of the characteristic polynomial

$$\det(\lambda \underline{I} - \underline{G}) = \lambda^n + \sum_{i=1}^{n} \lambda^{n-i} \phi_i . \tag{6.18}$$

(ii) Either the rank of $\underline{G}$ is n, implying that $\underline{G}$ is non-singular, or the rank of $\underline{G}$ is n-1, and $\underline{G}$ is singular, according to whether $\phi_n = 0$ or not.

Define s to be the largest integer such that $\phi_s \neq 0$, so that s = n in the non-singular case, and denote

$$\Phi(z) = z^s + \phi_1 z^{s-1} + \ldots + \phi_s . \tag{6.19}$$

Then in both cases

$$\underline{G}^{r+1} \Phi(\underline{G}) = \underline{0}_{nxn} \tag{6.20}$$

where r = 0 is the non-singular case, and r = n-s-1 in the singular; in the latter case note that

$$\underline{G}^r \Phi(\underline{G}) \neq \underline{0}_{nxn} . \tag{6.21}$$

When it is positive we refer to r as the system shift.

This links up to our earlier definitions as follows

Lemma 6.9

If $\underline{G}$ is singular then $\{\underline{F},\underline{G}\}$ has horizon $\leq n-1$. If $\underline{G}$ is non-singular then $\underline{G}$ has horizon $\leq n$. In particular if $\{F,G\}$ is observable then the inequalities are replaced by equalities.

Proof

We shall prove a slightly stronger result: Suppose that $\underline{G}$ has minimal polynomial

$$\lambda^d + \sum_{i=1}^{d} \lambda^{d-i} \alpha_i$$

then $\underline{G}$ is singular if and only if $\alpha_d \equiv 0$, in which case $\{\underline{G}^{d-1}, \ldots \underline{G}\}$ are linearly independent (ℓ.i) with $\{\underline{G}^d \ldots \underline{G}\}$ dependent (ℓ.d) so that $\underline{G}$ has horizon $\leq d-1$. If $\underline{G}$ is non-singular $\alpha_d \neq 0$, so that $\{\underline{G}^{d-1} \ldots I\}$ are ℓ.i with $\{\underline{G}^d \ldots \underline{I}\}\ell$.d, which is true if and only if $\{\underline{G}^d \ldots \underline{G}\}$ is ℓ.i and $\{\underline{G}^{d+1} \ldots G\}\ell$.d. Consequently we have proved the first part of the assertion, with $d = n$ in our case. The second follows on relacing $\underline{G}^i$ by $\underline{F}\,\underline{G}^i$ in the above argument.

6.3 Equivalence Results for the General Model-Time Invariant, Steady State

In this section we show that the forecast function for the DLM (6.1), (6.2) is identical to that for an ARIMA process, subject to certain conditions to be determined. In what follows we assume that $\underline{A}_t \equiv \underline{A}$ for all t, for which sufficient conditions are given in Chapter 4, and also that $\underline{G}$ is non-derogatory, so that it has minimal polynomial

defined above. We are then able to express the forecast function $\{y_t(k) : k \geq 1\}$ of the DLM in an interesting form. Godolphin and Harrison (1975), Godolphin and Stone (1980) considered polynomial-projecting predictors and looked at conditions under which DLMs have such predictors. Stone (1982) extended the results to forecast functions which are a polynomial after a certain lead time and also briefly considered the equivalence between DLMs and certain ARIMA(p,d,q) models. This section considers the completely general case and encompasses the results of the previous papers. The results contained in Theorems (6.12) to (6.14) have been found by Godolphin and Stone (private communication), however independent proofs are given here since they are needed for the time-varying results of Section 6.4.

We prefer to work with (6.1), (6.2) rather than

$$y_t = \underline{F}\,\underline{\theta}_t \tag{6.22}$$

$$\underline{\theta}_t = \underline{G}\,\underline{\theta}_{t-1} + \underline{W}_t \tag{6.23}$$

not only because most DLMs are expressed in the form (6.1), (6.2) in a natural way which is open to interpretability, but also because by re-expressing models in the form (6.22), (6.23) does not preserve observability. For example in the univariate case by defining $\underline{\phi}_t = \begin{pmatrix} \underline{\theta}_t \\ V_t \end{pmatrix}$ we can rewrite (6.1), (6.2) in the form

$$y_t = (\underline{F} \quad 1)\underline{\phi}_t \tag{6.24}$$

$$\underline{\phi}_t = \begin{pmatrix} \underline{G} & 0 \\ 0 & 0 \end{pmatrix} \underline{\phi}_{t-1} + \begin{pmatrix} \underline{W}_t \\ V_t \end{pmatrix} \tag{6.25}$$

then we have

Lemma 6.10

If ($\underline{F}$,$\underline{G}$) is observable then the system (6.24), (6.25) is observable if and only if $\underline{G}$ is non-singular.

Proof

The observability matrix of the extended system has determinant

$$\begin{vmatrix} \underline{F} & 1 \\ \underline{F}\,\underline{G} & 0 \\ \vdots & \vdots \\ \underline{F}\,\underline{G}^{n} & 0 \end{vmatrix} = \begin{vmatrix} \underline{F} \\ \vdots \\ \underline{F}\,\underline{G}^{n-1} \end{vmatrix} |\underline{G}|$$

which is non-zero if and only if $|\underline{G}|$ is non-singular.

We shall see later that singular $\underline{G}$ matrices play an important role in applications so that to represent the model in the form (6.22), (6.23) would mean that the new system would be unobservable, consequently the item of interest to us, namely the observable subsystem, has to be derived via the decomposition theorem.

Example 6.11

We shall see later that the Godolphin and Stone model

$$y_t = \theta_{1t} + \theta_{2t} + v_t$$
$$\theta_{1t} = \theta_{1t-1} + w_{1t}$$
$$\theta_{2t} = \theta_{1t-1} + w_{2t}$$

(ibid) is an example of a steady model. The augmented state representation (6.24), (6.25)is unobservable since $\underline{G}$ is singular. An observable version in the form (6.22-6.23) is

$$y_t = \theta_{1t} + \theta^*_{2t}$$
$$\theta_{1t} = \theta_{1t-1} + w_{1t}$$
$$\theta^*_{2t} = \theta_{1t-1} + w^*_{2t}$$

where θ^*_{2t} is defined to be $\theta_{2t} + v_t$, $w^*_{2t} = v_t + w_{2t}$.

We now develop the main theorems of equivalence where we assume that $\underline{F}$, $\underline{G}$ is observable.

Theorem 6.12

(i) For all $k > r+s$, $y_t(k) + \sum_{j=1}^{s} \phi_j y_t(k-j) = 0$ (6.26)

(ii) For $1 \leq k \leq r+s$, the forecast vector $\underline{f}_t$, given by $\underline{f}_t = \{y_t(1) \ \ldots \ y_t(r+s)\}^T$ has the updating equation

$$\underline{f}_t = \underline{G}^*_{r+s} \underline{f}_{t-1} + \underline{\alpha} e_t \tag{6.27}$$

where $\alpha_k = \underline{F}\underline{G}^k\underline{A}$, $\underline{\alpha} = (\alpha_1 \ \ldots \ \alpha_{r+s})^T$, $e_t = y_t - y_{t-1}(1)$

$$\underline{G}^*_{r+s} = \begin{pmatrix} \underline{0}_{r+s-1 \times 1} & \underline{I}_{r+s-1} \\ \hline \underline{0}_{1 \times r} & -\underline{\phi}_s^T \end{pmatrix} \quad r > 0, \quad \underline{G}_{0+s} = \begin{pmatrix} \underline{0}_{s-1 \times 1} & \underline{I}_{s-1} \\ \hline & -\underline{\phi}_s^T \end{pmatrix} \tag{6.28}$$

with $\underline{\phi}_s^T = (\phi_s, \phi_{s-1}, \ \ldots \ \phi_1)$.

Proof

Premultiplying equation (6.20) by $\underline{F}\underline{G}^{(k-r-s-1)}$ and postmultiplying by $\underline{m}_t$ yields (6.26). Premultiplying the Kalman updating formula (4.9) by $\underline{F}\underline{G}^k$, $1 \leq k \leq r+s$ gives

$$y_t(k) = y_{t-1}(k-1) + \alpha_k e_t \qquad (k=1, \ \ldots \ r+s-1) \tag{6.29}$$

and $y_t(r+s) = \underline{F}\underline{G}^{r+s+1} \hat{\underline{\theta}}_{t-1} + \alpha_{r+s} e_t$

$$= -\sum_{j=1}^{s} \phi_{s+1-j} \, y_{t-1}(r+j) + \alpha_{r+s} e_t \tag{6.30}$$

from (6.20). Equations (6.29), (6.30) are equivalent to the matrix equation (6.27) for the vector $\underline{f}_t$.

Remark

Equation (6.27) is of a similar form to the Kalman updating equation (4.9), however in general the equations are not the same since the matrices $\underline{G}^*$ and $\underline{G}$ can differ. In the singular case $\underline{G}^*_{r+s}$ has order one less than $\underline{G}^*_n$, namely

n-1, whatever the value of r. In the non-singular case $\underline{G}^*_{r+s} = \underline{G}^*_{0+s} = \underline{G}^*_n$, so $\underline{G}^*_{0+s}$ is similar to $\underline{G}$. Note that $\underline{G}^*_{r+s}$ is singular in the case of positive system shift ($r > 0$) with rank $(\underline{G}^*_{r+s}) = r+s-1$, otherwise $\underline{G}^*_{0+s}$ is non singular irrespective of the rank of $\underline{G}$.

The following three properties of $\underline{G}^*$ follow easily

Lemma 6.13

(1) The minimal polynomial of $\underline{G}^*$ is $z^r\Phi(z) \equiv \sum_{i=o}^{s} \phi_i z^{r+s-i}$

and so $(\underline{G}^*_{r+s})^r \Phi(\underline{G}^*_{r+s}) = \underline{0}_{r+s \times r+s}$ (6.31)

(2) $\underline{\kappa}\,\underline{G}^{*i} = (\underline{0}_{1xi}, 1, \underline{0}_{r+s-i-1})$ where $\underline{\kappa} = (1\ ,\ \underline{0}_{1xr+s-1})$ (6.32)

implying $\underline{\kappa}\,\underline{G}^{*i}\,\underline{\alpha} = \alpha_{i+1} \qquad i \leq r+s-1$ (6.33)

(3) $\underline{f}_t = \underline{G}^{*k} B^k \underline{f}_t + \sum_{j=0}^{k-1} (\underline{G}^* B)^j \underline{\alpha}\, e_t$, B the backward shift operator

$$= \underline{G}^{*k} \underline{f}_{t-k} + \sum_{j=0}^{k-1} \underline{G}^{*j} \underline{\alpha}\, e_{t-j}. \qquad (6.34)$$

Proof

(1) follows from the fact that $\underline{G}^*$ is a companion matrix whilst (3) is immediate from (6.27). (2) is proven by induction using

$$(\underline{0}_{1xj}, 1, \underline{0}_{1xr+s-j-1}) \begin{pmatrix} \underline{0}_{s-1x1} & I_{s-1} \\ - - - & - - - \\ -\underline{0}_{1xr} & -\underline{\phi}_s^T \end{pmatrix} = 1 . [\underline{G}^*]_j$$

$$= (\underline{0}_{1xj+1}, 1, \underline{0}_{1xr+s-j-2}).$$

We are now in a position to derive the main equivalence theorems for DLMs in the steady state. For the purposes of

the next theorem only we assume that the zeros of $\Phi(z)$ lie strictly inside the unit circle.

Theorem 6.14

For each $k = 1,2 \ldots$ the k step ahead predictor for the DLM coincides with k step ahead predictor for the autoregressive moving average (ARMA) model of order (s,r+s) given by

$$y_t + \phi_1 y_{t-1} + \ldots + \phi_s y_{t-s} = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_{r+s} \varepsilon_{t-r-s} \tag{6.35}$$

where

$$\beta_j = \sum_{i=0}^{j-1} \phi_i \alpha_{j-i} + \phi_j \ (1 \leq j \leq s), \beta_{s+j} = \sum_{i=0}^{s} \phi_i \alpha_{s+j-i} \quad (1 \leq j \leq r) \tag{6.36}$$

and the sequence $\{\varepsilon_t, t = 0 \pm 1, \ldots\}$ consists of purely random variables.

Proof

Let P(k) denote the statement of the theorem and consider the proof of P(1). Now

$$y_t(1) + \sum_{j=1}^{s} \phi_j y_{t-j}(1) \equiv \underset{\sim}{\kappa} \left(\underline{f}_t + \sum_{j=1}^{s} \phi_j \underline{f}_{t-j} \right). \tag{6.37}$$

Defining $\phi_o = 1$ and using (6.34), gives the right hand side of (6.37) as

$$\underset{\sim}{\kappa} \left\{ \sum_{j=0}^{s} \phi_j (\underline{G}^*_{r+s})^{r+s} \underline{f}_{t-r-s} + \sum_{j=0}^{s} \phi_j \sum_{i=1}^{r+s-j-1} (G^*_{r+s})^i \underline{\alpha} \, e_{t-i-j} \right\}$$

$$= \sum_{j=0}^{s} \phi_j \sum_{i=0}^{r+s-j-1} \alpha_{i+1} \, e_{t-i-j} \quad \text{using} \quad (6.31) \text{ and } (6.33)$$

$$= \sum_{j=0}^{r+s-1} b_{j+1} \, e_{t-j} \tag{6.38}$$

where $b_j = \sum_{j=0}^{j-1} \phi_i \alpha_{j-i} (1 \leq j \leq s)$, $b_{s+j} = \sum_{i=0}^{s} \phi_i \alpha_{s+j-i} \ (1 \leq j \leq r)$. (6.39)

On using the fact that $y_{t-j}(l) = y_{t+l-j} - e_{t+l-j}$ $(l \leq j \leq s)$ in equations (6.37) and (6.38) gives

$$y_t(l) + \sum_{j=0}^{s-l} \phi_{j+l} y_{t-j} = \sum_{j=0}^{r+s-l} \beta_{j+l} e_{t-j} \qquad (6.40)$$

where $\beta_j = b_j + \phi_j$, $(l \leq j \leq s)$, $\beta_{s+j} = b_{s+j}$ $(l \leq j \leq r)$ which proves P(l) since (6.40) is the defining relation for the one step ahead predictor for the ARMA model (6.35). That is if observations are generated by (6.35), then (6.40) is the conditional expectation of y_{t+l} given y_t with

$$e_{t-j} = E[\varepsilon_{t-j} | y^t] = y_{t+l-j} - y_{t-j}(l). \qquad (6.41)$$

Let k be an integer in the range $2 \leq k \leq s$. Now

$$(\phi_{k-1}, \cdots, \phi_o, 0_{1xr+s-k}) = \sum_{j=0}^{k-1} \phi_j \underset{\sim}{\kappa} (\underline{G}^*_{r+s})^{k-1-j}$$

using (6.32). Consequently

$$\sum_{j=0}^{k-1} \phi_j y_t(k-j) + \sum_{j=0}^{s-k} \phi_{k+j} y_{t-j-1}(1) \qquad (6.42)$$

$$= \sum_{j=0}^{k-1} \phi_j \underset{\sim}{\kappa} (\underline{G}^*_{r+s})^{k-1-j} \underline{f}_t + \sum_{j=0}^{s-k} \phi_{k+j} \underset{\sim}{\kappa} \underline{f}_{t-j-1}$$

$$= \underset{\sim}{\kappa} \Big\{ \sum_{j=0}^{k-1} \phi_j \underline{G}^{*k-1-j} \Big[(\underline{G}^*_{r+s})^{r+s-k+1} \underline{f}_{t-r-s+k+1} + \sum_{i=0}^{r+s-k} (\underline{G}^*_{r+s})^i \underline{\alpha} e_{t-i} \Big]$$

$$+ \sum_{j=0}^{s-k} \phi_{k+j} \Big[(\underline{G}^*_{r+s})^{r+s-j-k} \underline{f}_{t-r-s+k-1} + \sum_{i=0}^{r+s-j-k-1} (\underline{G}^*)^i \underline{\alpha} e_{t-j-1-i} \Big] \Big\}$$

using (6.34)

$$= \underset{\sim}{\kappa} \Big\{ \sum_{j=0}^{s} \phi_j (\underline{G}^*)^{r+s-j} \underline{f}_{t-r-s+k-1} + \sum_{j=0}^{k-1} \phi_j \sum_{i=0}^{r+s-k} (\underline{G}^*_{r+s})^{k-1-j+i} \underline{\alpha} e_{t-i}$$

$$+ \sum_{j=k}^{s} \phi_j \sum_{i=0}^{r+s-j-1} (\underline{G}^*)^i e_{t+k-j-i-1} \Big\}$$

$$= \underset{\sim}{\kappa} \left\{ (\underline{G}^*_{r+s})^r \Phi(\underline{G}^*_{r+s}) \underline{f}_{t+k-r-s-1} + \sum_{j=0}^{k} \phi_j \sum_{i=0}^{r+s-k} \alpha_{k+i-j} e_{t-i} \right.$$

$$\left. + \sum_{j=k}^{s} \phi_j \sum_{i=0}^{r+s-j-1} \alpha_{i+1} e_{t+k-j-1} \right\}$$

using (6.33), which by (6.31) and the definitions of (6.39)

$$= \sum_{j=0}^{r+s-k} b_{j+k} e_{t-j}. \qquad (6.43)$$

On substituting for $y_{t-j-1}(1)$ from (6.41) in (6.42) - (6.43) gives

$$\sum_{j=0}^{k-1} \phi_j y_t(k-j) + \sum_{j=0}^{s-k} \phi_{k+j} y_{t-j} = \sum_{j=0}^{r+s-k} \beta_{k+j} e_{t-j} \qquad (2 \leq k \leq j). \qquad (6.44)$$

But this is exactly the predictor updating equation for the ARMA model (6.35) obtained by replacing t by t+k and taking expectations conditional upon the data up to time t, so we have now proved P(k) for $1 \leq k \leq s$. It only remains to prove the result for $s < k \leq r+s$, since for $k > r+s$ the result is a consequence of (6.26). The proof for $s < k \leq r+s$ is similar to the above and is obtained by considering

$$\sum_{j=0}^{s} \phi_j y_t(k-j) = \underset{\sim}{\kappa} \sum_{j=0}^{s} \phi_j (\underline{G}^*_{r+s})^{k-1-j} \underline{f}_t$$

$$= \sum_{j=0}^{s} \phi_j \sum_{i=0}^{r+s-k} \alpha_{k+i-j} e_{t-i} = \sum_{i=0}^{r+s-k} \beta_{k+i} e_{t-i}.$$

Thus P(k) is proved for all $k \geq 1$, and so predictor equivalence is established.

Remark

We have not used the assumption that the zeros of $\Phi(z)$ lie inside the unit circle, and so we have the following important corollary.

Corollary 6.15

For each $k = 1,2 \ldots$ the k step ahead predictor for the DLM (6.1), (6.2) coincides with the k-step ahead predictor of the linear difference equation

$$y_t + \phi_1 y_{t-1} \cdots + \phi_s y_{t-s} = \varepsilon_t + \beta_1 \varepsilon_{t-1} \cdots + \beta_{r+s} \varepsilon_{t-r-s} \qquad (6.45)$$

where the β are as in (6.36), the sequence $\{\varepsilon_t\}$ are purely random variables, and the forecasts of (6.45) are obtained by taking conditional expectations using

$$\text{(a)} \quad E[\, y_{t+j} | y^t \,] = y_t(j) \qquad j>0$$

$$\text{(b)} \quad E[\, y_{t-j} | y^t \,] = y_{t-j} \qquad j \geq 0$$

$$\text{(c)} \quad E[\, \varepsilon_{t+j} | y^t \,] = 0 \qquad j>0$$

$$\text{(d)} \quad E[\, \varepsilon_{t-j} | y^t \,] = y_{t-j} - y_{t-j-1}(1) \qquad j \geq 0.$$

This is a very important result since it includes non-stationary models. In particular the zero's of $\Phi(z)$ may have zeros on the unit circle, in which case (6.45) may represent an ARIMA process, and quite generally (6.45) includes the general multiplicative seasonal model. For example if several of the ϕ's are zero then (6.45) may describe a non-multiplicative seasonal model.

We have made the assumption that $\underline{F}, \underline{G}$ is observable, in which case the matrix $\underline{G}$ is non-derogatory and (6.45) has its autoregressive part of lowest order, however the proof of Corollary 6.15 rests on Theorem (6.12), which is true for general $\underline{F}, \underline{G}$ provided that we define r,s and $\underline{\phi}_s$ to be the smallest values of r,s such that

$\underline{\phi}_s \neq 0$ and

$$\underline{F}\,\underline{G}^{r+1}\,\Phi(\underline{G}) = 0 \qquad (6.46)$$

if $\underline{G}$ is singular with $\Phi(z)$ given by (6.19); and $r = 0$, s the smallest integer such that

$$\underline{F}\,\Phi(\underline{G}) = 0 \qquad (6.47)$$

in the non-singular case. Then we have

Corollary 6.16

With $r, s, \underline{\phi}_s$ defined via (6.46), (6.47), Corollary 6.15 is true for any n-vector $\underline{F}$ and nxn matrix $\underline{G}$.

Remark

The order of the right hand side of (6.45) can be greater than, equal to or less than that of the left hand side. For if $r>0$

$$\beta_{r+s} = \sum_{i=0}^{s} \phi_i \underline{F}\,\underline{G}^{r+s-i}\,\underline{A} = \underline{F}\,\underline{G}^{r}\Phi(\underline{G})\underline{A}.$$

Now $\underline{A} \propto \underline{P}\underline{F}^T = \mathrm{cov}(\underline{\theta}_t | y^{t-1})\underline{F}^T$ which is in general non-zero so that from (6.21) $\beta_{r+s} \neq 0$ in general, so that the right hand side is of order r+s. If $r = 0$ then

$$\beta_j = \underline{F}\underline{G}(\underline{G}^{j-1} + \phi_1\underline{G}^{j-2} \ldots + \phi_{j-2}I)\underline{A} + \underline{\phi}_j \qquad 1 \leq j \leq s$$

which could be zero if the ϕ's are suitably defined, and the right hand side can be any order form zero to s. Note that if $\underline{A} \equiv 0$ then (6.45) becomes

$$\Phi(B)y_t = \Phi(B)\varepsilon_t$$

where B is the backward shift operator.

The parameters β_j of equation (6.45) or equivalently of an ARIMA type model are related to stability requirement by

Theorem 6.17

The stability conditions for the DLM (6.1), (6.2) coincide with the invertibility conditions for the process (6.45).

Proof

Taking z-transforms of (6.37), (6.38), and denoting the z-transorms of $y_t(1), y_t$, and e_t by $F_1(z)$, $y(z)$ and $E(z)$ then

$$F_1(z)\left[1 + \frac{\phi_1}{z^1} + \ldots + \frac{\phi_s}{z^s}\right] = E(z)\left[b_1 + \frac{b_2}{z} \ldots + \frac{b_{r+s}}{z^{r+s-1}}\right]$$

where we have ignored the transient terms, which we can do by assuming an infinite past history or equivalently zero starting conditions.

Now $$e_t = y_t - y_{t-1}(1)$$

so $$E(z) = Y(z) - \frac{F_1(z)}{z}$$

thus $$\frac{F_1(z)}{Y(z)} = \left(\sum_{j=1}^{r+s} b_j z^{r+s+1-j}\right) \div \beta(z)$$

where $$\beta(z) = z^{r+s}\left\{\frac{\Phi(z)}{z^s} + \sum_{j=1}^{r+s} b_j z^{-j}\right\} = \sum_{j=0}^{r+s} \beta_j z^{r+s-j}.$$

We can also show that $F_i(z)/Y(z)$ has denominator $\beta(z)$, where $F_i(z)$ is the z-transform of $y_t(i)$, i=2 ... r+s. The above recursions are stable if the zeros of $\beta(z)$ are less than one in modulus (Theorem 3.13) which is the invertibility condition for (6.45) and the theorem is proved.

If the DLM is uniformly completely observable and uniformly completely controllable then from Theorem 4.5 the linear discrete filter is uniformly asymptotically stable, that is the homogeneous part of the filter

$$\underline{\theta}_t = (I - \underline{A}_t \underline{F}_t)\, \underline{G}_t\, \underline{\theta}_{t-1} + \underline{A}_t y_t .$$

If we consider the time invariant case, then we only need to consider the steady state case, which exists by Corollary 4.6. In such a case the filter is also b.i.b.o. stable, that is treating the y_t as inputs, from Example 3.15 the zeros of

$$H(z) = \{ zI - (I - \underline{A}\underline{F})\underline{G} \}^{-1} zA$$

are less than one in modulus, where $\Theta(z) = H(z)Y(z)$. But $\underline{F}_i(z) = \underline{F}\underline{G}^i\, \Theta(z) = \underline{F}\underline{G}^i H(z)Y(z)$, so that if the zeros of $H(z)$ have modulus less than one, so do those of $F_i(z)/Y(z)$. In other words we have proved the following theorem

Theorem 6.18

In the time-invariant case, if the DLM is observable and controllable then the Kalman filter is stable, and the equivalent ARIMA model is invertible.

6.4 Equivalence Theorems for Time Varying DLMs

Section 6.3 concerned theorems for constant (time-invariant) DLMs in the steady state, whereas in fact we can extend Theorems 6.14 and 6.15 to include much more general models. We shall consider first the case where $\underline{F}, \underline{G}$ are constant but the gain matrix $\underline{A}_t$ is not constant.

This can happen in two cases

(1) When $V = \text{var } v_t$, $\underline{W} = \text{var } \underline{w}_t$ are constant but the matrix $\underline{A}_t$ has not converged to its limiting value. In fact we could also work in the terms of Corollary 6.16 in which case such a limit need not exist.

(2) $V_t = \text{var } v_t$, $\underline{W}_t = \text{var } \underline{W}_t$ are not constant in time.

By defining $\alpha_{k,t} = \underline{F}\underline{G}^k\underline{A}_t$, then by proceeding exactly as for the proof of Theorem 6.12 we have

Theorem 6.19

(i) for all $k>r+s$, $y_t(k) + \sum_{j=1}^{s} \phi_j y_t(k-j) = 0$

(ii) for $1 \leq k \leq r+s$ the forecast vector $\underline{f}_t$ has the updating equation

$$\underline{f}_t = G^*_{r+s}\underline{f}_{t-1} + \underline{\alpha}_t e_t$$

where $$\underline{\alpha}_t = (\alpha_{1,t} \cdots \alpha_{r+s,t})^T.$$

This enables us to deduce

Lemma 6.20

$$\underline{f}_t = (\underline{G}^*_{r+s})^k f_{t-k} + \sum_{j=0}^{k-1} (\underline{G}^*_{r+s})^j \underline{\alpha}_{t-j} e_{t-j} \qquad (6.48)$$

which with the aid of Lemma (6.13) parts (i) and (ii) enables us to mirror the proof of Theorem 6.14 and its corollaries yielding

Theorem 6.21

For each $k = 1,2 \ldots$ the k step ahead predictor for the DLM of this section coincides with the k-step ahead predictor for the linear difference equation

$$y_t + \phi_1 y_{t-1} + \cdots + \phi_s y_{t-s} = \varepsilon_t + \beta_{1t-1}\varepsilon_{t-1} + \cdots + \beta_{r+s,t-r-s}\varepsilon_{t-r-s}$$

where the $\beta_{j,t-j}$ are given by

$$\beta_{j,t-j} = \sum_{i=0}^{j-1} \phi_i \alpha_{j-i,t-j} + \phi_j$$

(recall that $\phi_{s+j} = 0$, $j>0$).

Proof

The proof is exactly as in Theorem 6.14, except that (6.48) is used instead of (6.34)

Remark

In the difference equation (6.35), or (6.45) the ε_t are usually taken to be independently and indentically distributed random variables; however it should be noted that the predictors derived for such models are also the optimal predictors (in the sense of smallest mean square error, that is the predictor in the conditional expectation) if the ε_t are merely independent. In fact they are also optimal if the β's are allowed to evolve in time in such a way that β_j depends on t through t-j, as happens in this theorem. So that the prescription under Corollary 6.15 still applies. For the DLM's the ε_t are indeed independent, as is well known, and under normality assumptions they have variance $\underline{F}\,\underline{G}\,\underline{C}_{t-1}\,\underline{G}^T\underline{F}^T + \underline{F}\,\underline{W}_t\underline{F}^T + \underline{v}_t$ which is constant if the error variances $\underline{V}$ and $\underline{W}$ are constant and if $\underline{C}_t$ has converged to a limit.

The above theorems quantifies the change of information that occurs through time as our prior knowledge is superceded. That is, if we are in the 'sensible' situation of having an observable and controllable model, then $\underline{A}_t \to \underline{A}$ and so $\alpha_{it} \to \alpha_i$, so that our forecasts come from models with

constant auto-regressive parts (not necessarily stationary) whose moving average parts converge in the sense of the parameters to a single model.

Finally we consider the completely general case where $\underline{F}(t)$, $\underline{G}(t)$ are time varying matrices, $\underline{F}(t) \varepsilon \mathbb{R}^n$, $\underline{G}(t)$an nxn matrix. Therefore there is some number k, real numbers $\phi_i(t)$ not all zero such that

$$\underline{F}(t+k)\underline{G}(t+k) \ldots \underline{G}(t+1) + \sum_{i=1}^{k} \phi_i(t+k-i)\underline{F}(t+k-i) \prod_{j=0}^{k-i-1} \underline{G}(t+k-i-j)=0$$

which on using the definition of the transition matrix $\underline{\Phi}(i,j)$ given in Chapter 3 can be rewritten as

$$\underline{F}(t+k)\Phi(t+k,t) + \sum_{i=1}^{k} \phi_i(t+k-i)\underline{F}(t+k-i)\underline{\Phi}(t+k-i,t) = 0. \tag{6.49}$$

In order that our models have a constant horizon for all t we require that (6.49) is only true for k = n. In other words we require that the matrix given by

$$\underline{M}(t+n-1,t) = \sum_{i=t}^{t+n-1} \underline{\Phi}(i,t)\underline{F}(i)^T\underline{F}(i)\underline{\Phi}(i,t) > 0$$

or in other words is positive definite (see Section 3.5). But this condition is satisfied if the system is uniformly completely observable .

With such a condition satisfied, two possibilities arise according to whether $\phi_n(t) = 0$ or not. That is we can mimic the exposition of Section 6.2 and define r,s in the case of $\phi_n(t) \equiv 0$ by r+s = n-1 where s is the largest integer such that $\phi_s \neq 0$ and in (6.49)

$$\underline{F}(t+r+s+1)\underline{\Phi}(t+r+s+1,t) + \sum_{i=1}^{s} \phi_i(t+r+s+1-i)\underline{F}(t+r+s+1-i)\underline{\Phi}(t+r+s+1-i,t) = 0 \tag{6.50}$$

or if $\phi_n(t) \neq 0$, define r to be zero, s = n so that postmultiplying (6.49) (with k = n) by $\underline{G}(t)$ gives

$$\underline{F}(t+n)\underline{\Phi}(t+n,t-1) + \sum_{i=1}^{n} \phi_i(t+n-i)\underline{F}(t+n-i)\underline{\Phi}(t+n-i,t-1) = 0. \quad (6.51)$$

In order that our conditions are consistent we require

Condition

Either $\phi_n(t)$ is non-zero for all t, so s = n, r = 0, or else s, (the largest integer such that $\phi_s(t+n-s)$ is non zero in (6.49) with k = n) is a constant throughout time, r = n-s-1.

Obviously we then have a constant horizon r+s in both cases as in the time invariant case. We can then prove

Theorem 6.22

(i) For all $k>r+s$, $y_t(k) + \sum_{j=1}^{s} \phi_j(t+k-j)y_t(k-j) = 0$ (6.52)

(ii) For $1 \leq k \leq r+s$ the forecast vector $\underline{f}_t$ has the updating equation

$$\underline{f}_t = \underline{G}^*_{r+s}\, \underline{f}_{t-1} + \underline{\alpha}_t\, e_t \quad (6.53)$$

where $$\underline{G}^*_{r+s} = \begin{pmatrix} \underline{0}_{r+s-1 \times 1} & I_{r+s-1} \\ \hline \underline{0}_{1 \times r} & -\underline{\phi}_s^T(t+r+s) \end{pmatrix} \quad \text{if } r>0$$

$$\underline{G}^*_{o+s} = \begin{pmatrix} \underline{0}_{s-1 \times 1} & I_{s-1} \\ \hline & -\underline{\phi}_s{}^T(t+r+s) \end{pmatrix}$$

with $\underline{\alpha}_t = (\alpha_{1,t} \cdots \alpha_{r+s,t})^T$, $\alpha_{i,t} = \underline{F}(t+i)\underline{\Phi}(t+i,t)\underline{A}_t$ and

$\underline{\phi}_s(t+r+s)^T = (\phi_s(r+t), \phi_{s-1}(t+r+1), \ldots \phi_1(t+r+s-1))$.

Proof

$y_t(k) = \underline{F}(t+k)\underline{\Phi}(t+k,t)\underline{m}_t$. Now (6.50) and (6.51) can both be expressed in the form (6.50). Putting $j = k-(r+s+1)$ for $k>r+s+1$ gives for $t=t+j$

$$\underline{F}(t+k)\underline{\Phi}(t+k,t+j) + \sum_{i=1}^{s} \phi_i(t+k-i)\underline{F}(t+k-i)\underline{\Phi}(t+k-i,t+j) = 0.$$

Postmultiplying by $\underline{\Phi}(t+j,t)\underline{m}_t$ and using the fact that

$$\underline{\Phi}(t+k,t+j)\underline{\Phi}(t+j,t) = \underline{\Phi}(t+k,t)$$

gives us the statement of (6.52). Premultiplying the Kalman updating formula (4.9) by $\underline{F}(t+k)\underline{\Phi}(t+k,t)$ $1 \leq k \leq r+s$ gives

$$y_t(k) = y_{t-1}(k+1) + \alpha_{k,t}e_t \qquad (6.54)$$

and $y_t(r+s) = \underline{F}(t+r+s)\underline{\Phi}(t+r+s,t-1)\underline{m}_{t-1} + \alpha_{r+s,t}e_t$

which from (6.50)

$$= -\sum_{j=1}^{s} \phi_j(t+r+s-j)y_{t-1}(r+s+1-j) + \alpha_{r+s,t}e_t. \qquad (6.55)$$

Equations (6.54), (6.55) are equivalent to the matrix equation (6.53).

We can now prove

Theorem 6.23

The time dependent DLM model is predictor equivalent to the time dependent ARMA type model

$$y_t + \phi_1(t-1)y_{t-1} \ldots + \phi_s(t-s)y_{t-s} =$$

$$\varepsilon_t + \beta_1(t-1) \ldots + \varepsilon_{t-r-s}\beta_{r+s}(t-r-s) \qquad (6.56)$$

where $\beta_j = \sum_{i=0}^{j-1} \phi_i(t-i)\alpha_{j-i,t-j} + \phi_j(t-j) \qquad 1 \leq j \leq s$

$$\beta_{s+j} = \sum_{i=0}^{s} \phi_i(t-i)\alpha_{s+j-i,t-s-j} \qquad 1 \leq j \leq r$$

and the sequence $\{\varepsilon_t\}$ consists of zero mean, uncorrelated random variables.

Proof

The proof cannot be proved by analogy with Theorem 6.14 since there is no equivalent to (6.31), so that the proof loses much of its elegance. Instead we have to resort to expressing $y_t(j)$ in terms of $y_{t-i}(1)$ from (6.54), (6.55) to prove P(1), where P(1) is the equivalence of the forecast $y_t(1)$ for all t. That is, from (6.54)

$$y_t(k+1) = y_{t+1}(k) - \alpha_{k,t+1}e_{t+1} \qquad 1 \leq k \leq r+s$$

so

$$y_t(k+1) = y_{t+k}(1) - \alpha_{k,t+1}e_{t+1} -$$

$$\alpha_{k-1,t+2}e_{t+2} \cdots - \alpha_{1,t+k}e_{t+k}.$$

Substituting into (6.55) gives

$$y_{t+r+s-1}(1) - \alpha_{r+s-1,t+1}e_{t+1} - \alpha_{r+s-2,t+2}e_{t+2} \cdots -$$

$$- \alpha_{1,t+r+s-1}e_{t+r+s-1}$$

$$= \alpha_{r+s,t}e_t - \sum_{j=1}^{s} \phi_j(t+r+s-j)[y_{t+r+s-j}(1) - \alpha_{r+s-j,t}e_t \cdots$$

$$- \alpha_{1,t+r+s-j-1}e_{t+r+s-j-1}]$$

Putting $t+r+s = t$ and defining $\phi_0(t) \equiv 1$ gives

$$y_{t-1}(1) + \sum_{j=1}^{s} \phi_j(t-j) y_{t-j-1}(1)$$

$$= \sum_{j=0}^{s} \phi_j(t-j) \sum_{i=1}^{r+s-j} \alpha_{i,t-i-j} e_{t-i-j} = \sum_{j=1}^{r+s} b_j(t-j) e_{t-j}$$

where $b_j(t-j) = \sum_{i=0}^{j-1} \phi_i(t-i) \alpha_{j-i,t-j}$

and since $y_{t-j-1}(1) = y_{t-j} - e_{t-j}$ then

$$y_{t-1}(1) + \sum_{j=1}^{s} \phi_j(t-j) y_{t-j} = \sum_{j=1}^{r+s} \beta_j(t-j) e_{t-j} \qquad (6.57)$$

where $\beta_j(t-j) = b_j(t-j) + \phi_j(t-j)$, which since $\phi_{s+j} = 0$ $j \geq 0$ is the same as in (6.56). But (6.57) is the one step-ahead predictor equation for (6.56) and so P(1) is proved. By a similar argument we can show that

$$y_t(k) + \sum_{j=1}^{k-1} \phi_j(t+k-j) y_t(k-j) + \sum_{j=k}^{s} \phi_j(t+k-j) y_{t+k-j}$$

$$= \sum_{j=k}^{r+s} \phi_j(t+k-j) e_{t+k-j} \qquad 2 \leq k \leq r+s$$

which is the defining relation for the k-step ahead predictor for (6.56). This together with Theorem 6.22 proves the result.

Remarks

(1) If $\phi_i(t)$ is independent of t for all i then the theorem reduces to Theorem 6.21.

(2) The theorem assumes knowledge of $\underline{F}(t+k), \underline{G}(t+k)$ for all t and k, since knowledge of $\phi_i(t+k-i)$ is needed. If this is known only up to some fixed k for each t, then

the theorem holds for all lead times less than or equal to k.

(3) There is an abuse of notation in the above proof since $\beta_j(t-j)$ depends on $\phi_1(t-1) \ldots \phi_{j-1}(t-j+1)$ so strictly speaking we should have $\beta_j(t-1)$. We prefer not to do this to preserve the identity of the $\beta_j(t-j)$s with those of Theorem 6.21 in the case where the ϕ_is are independent of t.

6.5 Covariance Properties

We now return to the time invariant case, where we assume that $\{\underline{F},\underline{G}\}$ is observable so that $\underline{G}$ is non-derogatory and the discussion of Section 6.2 applies. Using the model (6.1), (6.2) we have using (6.2) repeatedly that

$$\underline{\theta}_t = \underline{G}^k\underline{\theta}_{t-k} + \underline{G}^{k-1}\underline{w}_{t-k+1} + \underline{G}^{k-2}\underline{w}_{t-k+2} \ldots + \underline{w}_t. \qquad (6.58)$$

Using (6.58) to express $\underline{\theta}_{t-j}$ in terms of $\underline{\theta}_{t-n}$ and then using (6.1) gives

$$y_t + \phi_1 y_{t-1} \ldots + \phi_n y_{t-n} = \underline{F}(\underline{G}^n + \phi_1\underline{G}^{n-1} \ldots + \phi_n)\underline{\theta}_{t-n}$$

$$+ \underline{F}(\underline{G}^{n-1}\underline{w}_{t-n+1} + \underline{G}^{n-2}\underline{w}_{t-n+2} \ldots + \underline{w}_t)$$

$$+ \phi_1\underline{F}(\underline{G}^{n-2}\underline{w}_{t-n+1} \ldots + \underline{w}_{t-1}) + \ldots + \phi_{n-1}(\underline{F}\underline{w}_{t-n+1})$$

$$+ v_t + \phi_1 v_{t-1} \ldots + \phi_n v_{t-n}.$$

Using the fact that the minimal polynomial of $\underline{G}$ is given by 6.18 with $\phi_j = 0$ for j>s then we have on rearranging

Lemma 6.24

The DLM (6.1), (6.2) gives rise to the following relation

$$\sum_{i=0}^{s} \phi_i y_{t-i} = \sum_{i=0}^{s} \phi_i v_{t-i} + \sum_{i=0}^{n-1} \underline{\gamma}_i \underline{w}_{t-i}$$

with $\underline{\gamma}_i = \underline{F}\left(\sum_{j=0}^{i} \phi_j \underline{G}^{i-j}\right)$, $\phi_j = 0$ for $j>s$ (6.59).

In the terms of matrices (6.59) is

$$\begin{pmatrix} \underline{\gamma}_0 \\ \underline{\gamma}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underline{\gamma}_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \phi_1 & 1 & 0 \cdots & 0 \\ \phi_2 & \phi_1 & 1 & \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \phi_{n-1} & & & 1 \end{pmatrix} \begin{pmatrix} \underline{F} \\ \underline{F}\,\underline{G} \\ \cdot \\ \cdot \\ \cdot \\ \underline{F}\,\underline{G}^{n-1} \end{pmatrix}$$

or $\underline{\gamma} = \underline{\Phi}\,\underline{M}$

where $\underline{M}$ is the observability matrix. Consequently provided that the observability matrix has rank n, since $\underline{\Phi}$ and $\underline{M}$ are non-singular, knowledge of any two of $\underline{\gamma}$, $\underline{\Phi}$, $\underline{M}$ uniquely determines the third. In particular $\underline{\gamma}$ has rank n.

If we define the quantity z by the difference equation

$$z_t = \sum_{i=0}^{s} \phi_i y_{t-i}$$

then we have from Lemma 6.24,

Lemma 6.25

Provided that $E[v_t] = 0$, $E[\underline{w}_t] = 0$, $E[v_i v_j^T] = \delta_{ij} V$, $E[\underline{w}_i \underline{w}_j^T] = \delta_{ij} \underline{W}$ $E[v_i \underline{w}_j] = 0$ then z_t is a zero mean stationary stochastic process with autocovariance function given by

$$c_k = E[z_t z_{t+k}] = \underline{\gamma}_0 \underline{W} \underline{\gamma}_k^T + \underline{\gamma}_1 \underline{W} \underline{\gamma}_{k+1}^T \cdots + \underline{\gamma}_{n-k-1} \underline{W} \underline{\gamma}_{n-1}^T$$

$$+ V(\phi_k + \phi_1 \phi_{k+1} \cdots + \phi_{s-k} \phi_s)$$

$$= \sum_{i=0}^{n-k-1} \underline{\gamma}_i \underline{W} \underline{\gamma}_{i+k}^T + V \sum_{i=0}^{s-k} (\phi_i \phi_{i+k}) \qquad \text{for} \quad 0 \leq k \leq n-1 \qquad \phi_0 \equiv 1$$

$$c_n = V\phi_n$$

$$c_{n+k} = 0, \qquad k > 0. \qquad (6.60)$$

So if $\underline{G}$ is singular $\underline{c}_n = 0$, because $\phi_n \equiv 0$.

Since we are dealing with a stationary process we know from the Cramer-Wold factorisation that z_t has the same second order properties as

$$z_t \equiv \sum_{i=0}^{s} \phi_i y_{t-i} = \varepsilon_t + \bar{\beta}_1 \varepsilon_{t-1} \cdots + \bar{\beta}_{r+s} \varepsilon_{t-r+s}$$

where the $\{\varepsilon_t\}$ form an independent white noise sequence for some set $\{\bar{\beta}_i\}$. We are now in a position to relate our theorems of §6.3 to the covariance properties of the model.

Theorem 6.26

Provided that $\underline{P}$ is a solution to $\underline{P} = \underline{G}\,\underline{P}\,\underline{G}^T + \underline{W} - \dfrac{\underline{G}\underline{P}\underline{F}^T\underline{F}\underline{P}\underline{G}^T}{\underline{F}\underline{P}\underline{F}^T + v}$;

that is a steady state solution of the matrix Ricatti equation in the Kalman updating formulae, then the autocovariance function generated by (6.45), with β's defined by (6.36) where

$$\alpha_i = \underline{F}\underline{G}^i\underline{A} = \frac{\underline{F}\underline{G}^i\underline{P}\underline{F}^T}{\underline{F}\underline{P}\underline{F}^T + V}$$

is identical to that given by (6.60) in the sense of z_t, using the additional relation that

$$\text{var}(\varepsilon_t) = \sigma^2 = \underline{F}\underline{P}\underline{F}^T + V. \qquad (6.61)$$

Proof

(6.61) is a consequence of the remark of Section 6.4. The k^{th} covariance of z_t from (6.45) is

$$c_k = \sigma^2(\beta_k + \beta_1\beta_{k+1} \cdots + \beta_{s+r-k}\beta_{r+s}). \qquad (6.62)$$

From (6.60)

$$c_k = \sum_{i=0}^{n-k-1} \underline{\gamma}_i\underline{W}\underline{\gamma}_{i+k}^T + V\sum_{i=0}^{s-k} \phi_i\phi_{i+k}.$$

Now if the relation for P holds then

$$\underline{W} = \underline{P} - \underline{G}\underline{P}\underline{G}^T + \frac{\underline{G}\underline{P}\underline{F}^T\underline{F}\underline{P}\underline{G}^T}{\underline{F}\underline{P}\underline{F}^T + V}$$

$$\text{thus} \quad c_k = \sum_{i=0}^{n-k-1} \underline{\gamma}_i(\underline{P} - \underline{G}\underline{P}\underline{G}^T + \underline{G}\underline{A}\underline{F}\underline{P}\underline{G}^T)\underline{\gamma}_{i+k}^T + V\sum_{i=0}^{s-k} \phi_i\phi_{i+k}. \qquad (6.63)$$

Now from (6.59) $\underline{\gamma}_i\underline{G} = \underline{\gamma}_{i+1} - \underline{F}\phi_{i+1}$ which is true for $1 \leq i \leq n-1$ provided that we define $\gamma_n = \underline{F}(\sum_{j=0}^{n} \phi_j\underline{G}^{n-j}) = 0$.

$$\text{Thus} \sum_{i=0}^{n-k-1} \underline{\gamma}_i(\underline{P} - \underline{G}\underline{P}\underline{G}^T)\underline{\gamma}_{i+k}^T =$$

$$= \sum_{i=0}^{n-k-1} \{\underline{\gamma}_i\underline{P}\underline{\gamma}_{i+k}^T - (\underline{\gamma}_{i+1} - \underline{F}\phi_{i+1})\underline{P}(\underline{\gamma}_{i+k+1}^T - \phi_{i+k+1}\underline{F}^T)\}$$

$$= \sum_{i=0}^{n-k-1} (\underline{\gamma}_i \underline{P} \underline{\gamma}_{i+k}^T - \underline{\gamma}_{i+1} \underline{P} \underline{\gamma}_{i+k+1}^T + \underline{\gamma}_{i+1} \underline{P} \phi_{i+k+1} \underline{F}^T + \underline{F} \phi_{i+1} \underline{P} \underline{\gamma}_{i+k+1}$$

$$- \underline{F} \phi_{i+1} \underline{P} \phi_{i+k+1} \underline{F}^T)$$

$$= \underline{\gamma}_0 \underline{P} \underline{\gamma}_k^T - 0 + \sum_{i=0}^{n-k-1} (\phi_{i+k+1} \underline{\gamma}_{i+1} \underline{P} \underline{F}^T + \phi_{i+1} \underline{F} \underline{P} \underline{\gamma}_{i+k+1}$$

$$- \phi_{i+1} \phi_{i+k+1} \underline{F} \underline{P} \underline{F}^T). \qquad (6.64)$$

Now $\underline{\gamma}_0 = \underline{F}$, $\underline{P} \underline{F}^T = \underline{A} \sigma^2$, $\underline{F} \underline{P} \underline{F}^T = \sigma^2 - V$ and

$$\underline{\gamma}_j \underline{A} = \underline{F} \sum_{\ell=0}^{j} \phi_\ell \underline{G}^{\ell-j} \underline{A} = \beta_j - \phi_j + \underline{F} \phi_j \underline{A} = \beta_j - \phi_j (1 - \underline{F} \underline{A})$$

so $\gamma_j \underline{P} \underline{F}^T = \sigma^2 \underline{\gamma}_j \underline{A} = \beta_j \sigma^2 - \phi_j V$ since $\sigma^2 (1 - \underline{F} \underline{A}) = V$

thus (6.64) simplifies to

$$\sigma^2 \beta_k - \phi_k V + \sum_{i=0}^{n-k-1} [\, \phi_{i+k+1} (\beta_{i+1} \sigma^2 - \phi_{i+1} V)$$

$$+ \phi_{i+1} (\beta_{i+k+1} \sigma^2 - \phi_{i+k+1} V) - \phi_{i+1} \phi_{i+k+1} (\sigma^2 - V) \,]$$

$$= \sigma^2 \beta_k - \phi_k V + \sum_{i=0}^{n-k-1} [\, \phi_{i+k+1} \beta_{i+1} \sigma^2 + \phi_{i+1} \beta_{i+k+1} \sigma^2$$

$$- \phi_{i+1} \phi_{i+k+1} (\sigma^2 + V) \,]. \qquad (6.65)$$

The third term in (6.63) is $\sum \underline{\gamma}_i (\underline{G} \underline{A} \sigma^2 \underline{A}^T \underline{G}^T) \underline{\gamma}_{i+k}^T$ which from (6.59) and (6.36) is

$$\sum_{i=0}^{n-k-1} \sigma^2 (\beta_{i+1} - \phi_{i+1})(\beta_{i+k+1} - \phi_{i+k+1}) \qquad (6.66)$$

so that using (6.66) and (6.65) in (6.63) we have

$$c_k = \sigma^2\beta_k - \phi_k V + \sigma^2 \sum_{i=0}^{n-k-1} \beta_{i+1}\beta_{i+k+1} - \sum_{i=0}^{n-k-1} \phi_{i+1}\phi_{i+k+1} V$$

$$+ V \sum_{i=0}^{s-k} \phi_i \phi_{i+k}$$

which since $\phi_j = 0$, $j>s$ gives

$$c_k = \sigma^2 \sum_{i=0}^{n-k} \beta_i \beta_{i+k}. \qquad (6.67)$$

There are two cases to consider

(i) if $n = s$, so $r = 0$ then (6.67) is exactly (6.62)

(ii) $n = r+s+1$. Then in (6.66) we have defined β_n from (6.36) as

$$\beta_n = \phi_n + \sum_{i=0}^{n-1} \phi_i \underline{F}\underline{G}^{n-i}\underline{A} = \sum_{i=0}^{s} \phi_i \underline{F}\underline{G}^{1+r-i}\underline{A}$$

$$= \underline{G}^{r+1}\Phi(\underline{G})\underline{A} = 0$$

thus in both cases

$$c_k = \sigma^2 \sum_{i=0}^{r+s-k} \beta_i \beta_{i+k}$$

which is (6.62) and the result is proved.

So Theorem 6.26 assures us that in the steady state, the predictors of a time-invariant DLM are precisely those of an ARMA type model whose autoregressive parameters are obtained from the minimal polynomial of $\underline{G}$ and whose moving average parameters can be obtained from an examination of the autocovariance properties of the derived process z_t. Of course before we reach this state the discussion of Section 6.4 applies.

6.6 Structural Properties (of the general model)

We now illustrate some of the consequences of the above theorems for the structural properties of DLM's which can be used to describe various types of univariate time series. We assume that s is a fixed positive integer and begin by examining the dimension of $\underline{G}$. The case r>0 is the so-called forward shifted model, and it follows from the definition of r that $\underline{G}$ is singular in this case, that is n = r+s+1 whereas rank ($\underline{G}$) = r+s. On the other hand if r = 0 then there are two possibilities, n = s or n = s+1. The first corresponding to a non-singular system matrix $\underline{G}$ and the second to a singular system matrix. Consider the case n=s= rank ($\underline{G}$), then in the non-singular case the characteristic polynomial is given by (6.18). In such circumstances severe restrictions are placed upon values of $\beta_1 \ldots \beta_s$. For example

$$\beta_s = \phi_s + \sum_{j=0}^{s-1} \phi_j \alpha_{s-j} = \underline{F}\Phi(\underline{G})\underline{A} + \phi_s(1-\underline{F}\underline{A})$$

$$= \phi_s(1-\underline{F}\underline{A})$$

since $\Phi(\underline{G}) = \underline{0}_{nxn}$. However $1 - \underline{F}\underline{A} = \dfrac{\underline{V}}{\underline{F}\underline{P}\underline{F}^T + V}$, so $0 \leq 1 - \underline{F}\underline{A} \leq 1$

thus β_s has the same sign as ϕ_s and

$$0 \leq |\beta_s| \leq |\phi_s| . \qquad (6.68)$$

Strict inequalities hold provided that V is non-zero and the model is controllable. However the region (6.68) is to be compared with the invertibility/stability region for

the β's, which has $|\beta_s|<1$. In typical ARMA cases $|\phi_s|<1$, which supports the conjecture of Godolphin (1976) that DLM's with non-singular system matrices can be more severely restricted even than the polynomial DLM's discussed by Harrison and Stevens (1976).

It follows that if $s>0$ and $r\geq 0$ are given then the choice of $n = r+s+1$ is to be preferred on the grounds that if produces an observable model of smallest rank which does not impose the obvious restrictions that the case $n = r+s$ does. Such a choice has a slightly counter-intuitive appearance to it, since the dimension for the system vector $\underline{\theta}_t$ may appear to comprise one more component than might be expected. For example the case $r = 0$ is considered by Harrison and Stevens (1976), Smith (1981).

If we make the selection

$$\underline{F} = \underline{F}^*_{r+s+1} = (1, \underline{0}_{1 \times r+s})$$

together with the matrix

$$\underline{G} = \underline{G}^*_{r+s+1} = \begin{pmatrix} \underline{0}_{r+s \times 1} & I_{r+s} \\ \cdots & \cdots \\ \underline{0}_{1 \times r+1} & -\underline{\phi}_s^T \end{pmatrix}$$

then the model is observable since using (6.32) we have that $[F^T, (FG)^T, \ldots (FG^{r+s})^T]^T = I_{r+s+1}$ which has rank $r+s+1$. Consequently it follows from Lemma 6.4 (iii) that the model

$$\begin{aligned} \underline{F} &= \underline{F}^*_{r+s+1} \quad \theta \, (\underline{G}^*) \\ \underline{G} &= \underline{G}^* \end{aligned} \tag{6.69}$$

is observable provided that the polynomials $\Theta(z)$ and $z\Phi(z)$ are coprime. In general we know that for $\{\underline{F},\underline{G}\}$ to be observable $\underline{G}$ must be non-derogatory, consequently $\underline{G}$ is similar to $\underline{G}^*$, and so a general specification is provided by

$$\begin{aligned} \underline{F} &= \underline{F}^*_{r+s+1} \quad \Theta(\underline{G}^*)\underline{R}^{-1} \\ \underline{G} &= \underline{R}\,\underline{G}^*_{r+1+s}\,\underline{R}^{-1} \end{aligned} \tag{6.70}$$

for any non-singular matrix $\underline{R}$ of order $(r+1+s)x(r+s+1)$, which is strictly equivalent to (6.69).

The theorems have been mainly concerned with deriving and ARMA-type model from a given DLM. All such procedures involve the determination of the steady state Kalman gain vector A, which is a simple proposition numerically since we can use the Kalman filter. The procedure for reversing the process is in general more difficult. That in given r,s and a set of parameters $\beta_1 \ldots \beta_{r+s}$ we can invert (6.36) to obtain

$$\begin{pmatrix} 1 \\ \alpha_1 \\ \vdots \\ \alpha_{r+s} \end{pmatrix} = \begin{pmatrix} 1 & & \\ \theta_1 & 1 & \underline{0} \\ \theta_2 & \theta_1 & \\ \vdots & & \ddots \\ \theta_{r+s} & \theta_{r+s-1} & \ldots\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ \beta_1 \\ \\ \vdots \\ \beta_{r+s} \end{pmatrix}$$

where $\quad \theta_k + \phi_1\theta_{k-1} + \ldots + \phi_{k-1}\theta_1 + \phi_k = 0 \qquad 1 \leq k \leq s$

$$\theta_k + \phi_1\theta_{k-1} + \ldots + \phi_s\theta_{k-s} = 0 \qquad s+1 \leq k \leq r+s$$

and where $\phi_1 \ldots \phi_s$ are the coefficients of $\Phi(z)$.

Since the matrix is lower triangular, the α_i are uniquely determined. Having determined these, then

$$\underline{H}\,\underline{A} = \begin{pmatrix} \underline{F}\,\underline{G} \\ \vdots \\ \underline{F}\,\underline{G}^{r+s} \end{pmatrix} \quad \underline{A} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{r+s} \end{pmatrix}$$

$\underline{H}$ is the $(r+s) \times (n)$ matrix whose i^{th} row is $\underline{F}\,\underline{G}^i$. If $\underline{G}$ is non-singular then $r+s = n$, thus $\underline{H}$ is non-singular so that $\underline{A}$ is uniquely determined. In the singular case as already mentioned

$$\sigma^2 = \underline{F}\underline{P}\underline{F}^T + V \quad \text{or} \quad (1-\underline{F}\underline{A})\sigma^2 = V$$

thus

$$\underline{F}\underline{A} = (1-\sigma^2/V) \tag{6.71}$$

and since $\{F^T,(FG)^T \ldots,(FG^{r+s})^T\}^T$ is non-singular then $\underline{A}$ is uniquely determined given $1 - \sigma^2/V$.

In both these cases it is then required to solve for V and W from (6.71) and

$$\underline{A} = \frac{\underline{P}\underline{F}^T}{\underline{F}\underline{P}\underline{F}^T + V}$$

$$\underline{P} = \underline{G}\underline{P}\underline{G}^T + \underline{W} - \frac{\underline{G}\underline{P}\underline{F}^T\underline{F}\underline{P}\underline{G}^T}{\underline{F}\underline{P}\underline{F}^T + V}$$

Some conditions under which it will be possible to do this have already been given. This topic is continued in Chapter 7.

CHAPTER 7

INVERTIBILITY REGIONS AND PRACTICAL DLMS

7.1 Introduction

In this chapter we illustrate some of the results of Chapter 6 by applying them to typical DLMs. In particular we comment on the stability regions that are covered by DLMs which possess a certain intuitive appeal. The appropriate theory is developed in Section 2, and Sections 3 to 5 consists of examples. In §7.6 a slightly different way of writing some useful DLMs is given, and §7.7 shows how the Kalman filter can be applied to them. Section 7 also summarises the examples of the chapter concerning some frequently used ARIMA models and their DLM equivalents.

7.2 Theoretical Results

The specification of general $\underline{F}$ and $\underline{G}$ matrices in the time-invariant case for univariate observations is provided by (6.70). However in applications it will be necessary to specify V and the elements of the positive semi-definite matrix $\underline{W}$. In practical situations if we do not employ a statistical estimation procedure then Harrison and Stevens (1976), Godolphin and Stone (1980) and others remark that it is only the specification and updating of variances that can be carried out with any confidence. Consequently in practice we require $\underline{W}$ to be diagonal, or expressible in the form $\underline{B}\,\underline{W}_d\,\underline{B}^T$ where $\underline{B}$ is a fixed known

matrix and $\underline{W}_d$ diagonal, whose elements can vary. The latter case can be reduced to the former by using the equivalence between $\{\underline{F},\underline{G},\ \underline{B}\,\underline{W}_d\,\underline{B}^T\}$ and $\{\underline{F}\,\underline{B},\ \underline{B}^{-1}\underline{G}\,\underline{B}\ ,\ \underline{W}_d\}$. We therefore have two sets or regions

$$\underline{R}_I = \{V,\underline{W}\,|\,V\geq 0,\underline{W}\geq 0\} \tag{7.1}$$

$$\underline{R}_{II} = \{V,\underline{W}\,|\,V\geq 0,\underline{W}\geq 0 \text{ and diagonal}\} \tag{7.2}$$

where the second is of much more use practically. A general DLM is represented by (6.70) with $V,\underline{W}$ belonging to (7.1) or (7.2).

We know from §6.3 and in particular Theorem 6.14 and Corollary 6.15 that a time-invariant DLM is predictor equivalent to the moving average model

$$y_t + \phi_1 y_{t-1} + \ldots + \phi_s y_{t-s} = \varepsilon_t + \beta_1\varepsilon_{t-1} + \ldots + \beta_{r+s}\varepsilon_{t+r-s}$$

where the β_i are given by (6.36), namely

$$\begin{aligned} \beta_j &= \sum_{i=0}^{j-1} \phi_i \alpha_{j-i} + \phi_j \qquad 1\leq j\leq s \\ \beta_{s+j} &= \sum_{i=0}^{s} \phi_i \alpha_{s+j-i} \qquad 1\leq j\leq r. \end{aligned} \tag{7.3}$$

It is then natural to ask what is the nature of the stability/invertibility region under this equivalence for a DLM with $\{\underline{F},\underline{G}\}$ fixed but where $\{V,\underline{W}\}$ are allowed to range over R_I or R_{II}. For example we have already seen that too small a $\underline{G}$ matrix precludes us from covering the full region.

For R_I, the equivalence gives a mapping, dependent upon $\underline{F},\underline{G}$

$$R_I \xrightarrow{\underline{F},\ \underline{G}} \{\beta_1 \ \ldots \ \beta_{r+s}\} \qquad (7.4)$$

provided that a steady state $\underline{A}$ exists. First we shall show

Theorem 7.1

For fixed $\underline{G}$, the image of the mapping (7.4) is the same for all $\underline{F}$, provided that $\{\underline{F}, \underline{G}\}$ is observable.

Proof

Let $\underline{F},\underline{G}$ be any two matrices for which the observability matrix $\underline{M}$ has rank n. Then from Lemma 6.24 there is a non-singular matrix $\underline{\gamma}$, which depends upon $\underline{F}$, $\underline{G}$ given by

$$\underline{\gamma} = \underline{\Phi}\ \underline{M}$$

using the notation of the Lemma, where $\underline{\Phi}$ is a lower triangular matrix with 1's on the diagonal and $\Phi_{ij} = \phi_{j-i}$ for $j>i$. Consequently the DLM can be written in the form

$$\sum_{i=0}^{s} \phi_i y_{t-i} = \sum_{i=0}^{s} \phi_i v_{t-i} + \sum_{i=0}^{n-1} \underline{\gamma}_i \underline{w}_{t-i}$$

$$= \sum_{i=0}^{s} \phi_i v_{t-i} + (\underline{\gamma}\underline{w}_t)_1 + (\underline{\gamma}\underline{w}_{t-1})_2 + \ldots +$$

$$+ (\underline{\gamma}\underline{w}_{t-n+1})_n \qquad (7.5)$$

where the subscripts denote components of the column vector. Now from Theorem 6.26 the β_i's can be obtained from the autocovariance function of (7.5), but this function depends only on $\underline{\gamma}\,\underline{W}\,\underline{\gamma}^T$ and since the two sets $\{\underline{\gamma}\,\underline{W}\,\underline{\gamma}^T \mid \underline{W} \geq 0,\ \underline{\gamma}$ fixed and non-singular$\}$ and $\{\underline{W} \mid \underline{W} \geq 0\}$ are

identical, then the β_i's are independent of $\underline{\gamma}$ and hence $\underline{F}$. The theorem is therefore proved.

From the proof it follows that a 'canonical form' for the mapping (7.4) is (on putting $\underline{W}_t = \underline{\gamma}^{-1}\underline{W}_t$)

$$\sum_{j=0}^{s} \phi_j y_{t-j} = w_{1t} + w_{2t-1} + \ldots + w_{nt-n+1} + \sum_{j=0}^{s} \phi_j v_{t-j} \qquad (7.6)$$

where w_{it} is a scalar, $\underline{w}_t = \begin{pmatrix} w_{1t} \\ \cdot \\ \cdot \\ \cdot \\ w_{nt} \end{pmatrix}$,

$$E[\underline{w}_j \underline{w}_k]^T = \delta_{jk} \underline{W} \qquad (7.7)$$

and $\underline{W}$ is positive semi-definite.

This enables us to prove

Theorem 7.2

The mapping (7.4) under R_I allows the β_i to cover the full stability region, provided that n-1=r+s, i.e. n=r+s+1.

Proof

Put V = 0, $w_{it} = \beta_{i-1}\varepsilon_t$, then if n=r+s+1 (7.6) gives

$$\sum_{j=0}^{s} \phi_j y_{t-j} = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_{r+s}\varepsilon_{t-r-s}$$

and the matrix $\underline{W}$ of (7.6) is positive semi-definite as required. Now by varying the β's we can cover the full region, which completes the proof.

So under $\underline{R}_I$, DLMs encompass all the traditional linear difference models. Now consider the region R_{II} and the mapping

$$\underline{R}_{II} \longrightarrow \{\beta_i, \ldots \beta_{r+s}\} \qquad (7.8)$$

again under (7.3) with $\underline{F},\underline{G}$ fixed. It follows from the proof of Theorem 7.2 that a canonical form for the mapping (7.8) is given by (7.6) but where now instead of (7.7) for a fixed $\underline{B}$ $(\equiv\underline{R})$,

$$E[\underline{w}_j\underline{w}_k^T] = \delta_{jk}\underline{B} \begin{pmatrix} w_1 & & & \underline{0} \\ & w_2 & & \\ & & \ddots & \\ \underline{0} & & & w_n \end{pmatrix} \underline{B}^T \qquad (7.9)$$

The difference is because $\{\underline{R}\,\underline{W}\,\underline{R}^T | \underline{W}\geq 0$ and diagonal$\} \neq \{\underline{W} | \underline{W}\geq 0$ and diagonal$\}$. It will be algebraically more convenient to use the autocorrelation of the derived process z_t defined in Chapter 6 rather than the β_i's. Thus the mapping (7.8) can be obtained by looking at the mapping

$$\underline{R}_{II} \longrightarrow \{\rho_i, \ldots \rho_{r+s}\}$$

using (7.6) and (7.9) with $\underline{B}$ fixed. To summarise we have the following, which we formally state as a theorem:

Theorem 7.3

Let $\{y_t\}$ denote a DLM with canonical form given by (7.6), (7.9). Then the autocorrelation function for

$$z_t = \sum_{j=0}^{s} \phi_j y_{t-j} \text{ is}$$

$$\rho_k = \frac{E[z_t z_{t+k}]}{E[z_t^2]} = \frac{\sum_{i=1}^{n-k} \underline{b}_i \underline{W} \underline{b}_{i+k}^T + V \sum_{i=0}^{s-k} \phi_i \phi_{i+k}}{\sum_{i=1}^{n-k} \underline{b}_i \underline{W} \underline{b}_i^T + V \sum_{i=0}^{s} \phi_i^2} \qquad (7.10)$$

where $\underline{b}_i$ in the i^{th} row of the $\underline{B}$ matrix in (7.9), and $\underline{W} = \text{diag}\{W_1 \ldots W_n\}$.

The proof follows immediately from (7.6) and (7.9). So an examination of the mapping (7.8) is found by considering the mapping from

$$R_{III} \equiv \{V, W_1 \ldots W_n \mid V \geq 0, W_i \geq 0\} \longrightarrow \{\rho_1 \ldots \rho_{r+s}\} \qquad (7.11)$$

under (7.10).

Now $$\underline{b}_i \underline{W} \underline{b}_{i+k}^T = (b_{i1} \ldots b_{in}) \begin{pmatrix} W_1 & & \underline{0} \\ & \ddots & \\ \underline{0} & & W_n \end{pmatrix} \begin{pmatrix} b_{i+k\,1} \\ \vdots \\ b_{i+k\,n} \end{pmatrix}$$

$$= \sum_{j=1}^{n} b_{ij} W_j b_{i+k\,j}.$$

Thus we can write

$$\rho_k = \frac{\sum_{j=1}^{n} c_{k\,j} W_j + c_{k\,n+1} V}{\sum_{j=1}^{n} c_{0\,j} W_j + c_{0\,n+1} V} \qquad (7.12)$$

where
$$\left.\begin{aligned} c_{k\,j} &= \sum_{i=1}^{n-k} b_{i\,j} b_{i+k\,j} \qquad 1 \leq j \leq n \\ c_{k\,n+1} &= \sum_{i=0}^{s-k} \phi_i \phi_{i+k}. \end{aligned}\right\} 0 \leq k \leq r+s \qquad (7.13)$$

Since $\rho_0 = 1$, the image of the mapping (7.11) consists of a region in r+s dimensional space whose co-ordinate axes are $\rho_1, \ldots \rho_{r+s}$, that is we have a mapping from the set $R_{III} \varepsilon\ \mathbb{R}^{n+1} \longrightarrow \mathbb{R}^{r+s}$ given by

$$\underline{x} \longrightarrow \left(\frac{\underline{c}_1 \cdot \underline{x}}{\underline{c}_0 \cdot \underline{x}}, \frac{\underline{c}_2 \cdot \underline{x}}{\underline{c}_0 \cdot \underline{x}}, \ldots \frac{\underline{c}_{r+s} \cdot \underline{x}}{\underline{c}_0 \cdot \underline{x}} \right)^T \qquad (7.14)$$

where $\underline{c}_k \cdot \underline{x}$ is the dot product between $\underline{c}_k = (c_{k1} \cdots c_{kn+1})$ and $\underline{x} = (W_1 \ldots W_n \; V) \equiv (x_1 \ldots x_{n+1})^T$. Note that $c_{0j} > 0$ since the matrix $\underline{B}$ is non-singular, and without loss of generality we can put $x_{n+1} = V$ equal to 1, by say defining $W_j = W_j/V$.

Define the mapping $y_i = c_{0i}\, x_i \,/\, (c_{01}x_1 + \ldots + c_{0n}x_n + c_{0n+1})$ for i=1 ... n. This is clearly a one-one mapping on $\mathbb{R}^n$ because the denominator is non-zero and the inverse mapping is $x_j = y_j\, c_{0n+1} \,/\, (c_{0j}\{1-\Sigma y_i\})$, and further $0 \leq y_k \leq 1$ with $x_k \geq 0$ under the inverse mapping. Thus we can reparameterise (7.14) to yield the mapping from $C \subset \mathbb{R}^n \rightarrow \mathbb{R}^{r+s}$ given by

$$\left(\frac{c_{11}}{c_{01}} y_1 + \frac{c_{12}}{c_{02}} y_2 + \ldots + \frac{c_{1n}}{c_{0n}} y_n + \frac{c_{1\,n+1}}{c_{0n+1}} (1-y_1-..y_n) \;,\; \frac{c_{21}}{c_{01}} y_1 \ldots + \right.$$

$$\left. \frac{c_{2\,n+1}}{c_{0\,n+1}} (1-y_1- ..y_n), \ldots \;,\; \frac{c_{r+s\,1}}{c_{01}} y_1 + \cdots \frac{c_{r+s\,n+1}}{c_{0\,n+1}} (1-y_1-..y_n) \right)$$

where $y_i \geq 0$, $\Sigma\, y_i \leq 1$ and C is the region covered by $\{y_1 \cdots y_n\}$. If we now put $y_{n+1} = 1-y_1- \ldots -y_n$, then every point in the image is of the form $\Sigma\, y_i\, u_i$ where $y_i \geq 0$, $\Sigma\, y_i = 1$, and

$$u_i = \left(\frac{c_{1i}}{c_{0i}}, \ldots , \frac{c_{r+s\,i}}{c_{0\,i}} \right). \qquad (7.15)$$

Thus every point is a convex combination of the u_i, and so the image which consists of all such combinations is by definition a convex hull of the u_i. Now we can select a minimal set of the u_i's such that the image is the convex hull of these points, and it is easy to show that these points are vertices (see for example Trustram (1971)).

The above means that we can now fully characterise the image. That is

Theorem 7.4

The image under the mapping (7.8) from R_{II} to the autocorrelation space $\{\rho_1 \ldots \rho_{r+s}\}$ is the convex hull of its vertices, which are a subset of the points u_i of (7.15).

The theorem tells us that the image in the auto-correlation plane is bounded by vertices and straight lines. This means that if the invertibility region is not a convex polytope (the convex hull of a finite set) then it is impossible to cover the region under the mapping R_{II}. In fact the invertibility regions are convex sets with no vertices, however we can make them into convex polytopes by including the limit points to make the above statement non-trivial. The implications of this are that for higher order models, practical DLM's can only cover part of the invertibility region of the equivalent ARIMA model. This is illustrated in Section 4.

Note that the region may depend upon u_{n+1}

$$u_{n+1} = \left(\frac{c_{1\ n+1}}{c_{0\ n+1}} , \ldots , \frac{c_{r+s\ n+1}}{c_{0\ n+1}} \right)$$

which is soley a function of the ϕ's, so that if the region is to be independent of the ϕ's then in some sense the v_t term is redundant (see (7.13)).

The following sections contain examples which use the theory of both Chapter 6 and this section, so that the various theorems and results can be seen in context.

7.3 An Uncontrollable Model

Chapter 6 examined the role of observability, however controllability is not so easily dealt with. Although controllability is a sufficient condition for the convergence of the filter, it is not necessary, and DLM's need not be predictor equivalent to their controllable subsystems. Consider the model

Example 7.5

$$\begin{aligned} y_t &= (1 \;\; 1)\,\underline{\theta}_t \\ \underline{\theta}_t &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\underline{\theta}_{t-1} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} w_t \end{aligned} \qquad (7.16)$$

which is observable but clearly uncontrollable. The updates for $\underline{P}_t$ are given by (4.37) under the assumption that $w_t \sim N(0, W)$, which on writing $\underline{P}_{t-1} = \begin{pmatrix} P_1 & P_2 \\ P_2 & P_3 \end{pmatrix}$ and substituting for $\underline{F}$, $\underline{G}$, $\underline{W}$, and V (=0) gives

$$\underline{P}_t = \begin{pmatrix} P_1+2P_2+P_3+W & P_2+P_3 \\ P_2+P_3 & P_3 \end{pmatrix} - \begin{pmatrix} P_1+2P_2+P_3 \\ P_2+P_3 \end{pmatrix} \frac{(P_1+2P_2+P_3 \;\; P_2+P_3)}{P_1+2P_2+P_3}$$

so that $P_{1t} = W$, $P_{2t}=0$, $P_{3t} = (P_1P_3 - P_2^2)/(P_1+2P_2+P_3)$ and after one recursion $P_{3t} = (P_{3t-1}W)/(P_{3t-1} + W)$ which tends to the limit 0 as t increases.

The updates for the system parameter are

$$\underline{m}_t = \begin{pmatrix} m_{1t} \\ m_{2t} \end{pmatrix} = \begin{matrix} m_{1t-1} + m_{2t-1} + A_{1t}(y_t - m_{1t-1} - 2m_{2t-1}) \\ m_{2t-1} + A_{2t}(y_t - m_{1t-1} - 2m_{2t-1}) \end{matrix}$$

where $A_{1t} = (P_{1t} + P_{2t})/(P_{1t} + 2P_{2t} + P_{3t})$ with limit 1

and $A_{2t} = (P_{2t} + P_{3t})/(P_{1t} + 2P_{2t} + P_{3t})$ with limit 0.

The forecasts are $y_t(k) = \underline{F}\underline{G}^k\underline{m}_t = m_{1t} + (k+1)m_{2t} = y_t + km_{2t}$

therefore in the limit $m_{2t}=m_{2t-1}=\hat{m}$ say, and

$$y_t(k) = y_t + k\hat{m} \tag{7.17}$$

Now $\underline{G}$ has minimal polynomial $\underline{G}^2 - 2\underline{G} + \underline{I}$ so from Theorem 6.21, $\beta_{1t-1} = \alpha_{1t-1} + \phi_1 = \underline{F}\,\underline{G}\,\underline{A}_{t-1} - 2$

$$= \frac{-P_{1t-1} - P_{2t-1}}{P_{1t-1} + 2P_{2t-1} + P_{3t-1}} \tag{7.18}$$

and $\beta_{2t-2} = \alpha_{2t-2} + \phi_1\alpha_{1t-2} + \phi_2 = \underline{FG}^2\underline{A}_{t-2} - 2\underline{FGA}_{t-2} + 1 = 0$, consequently the DLM (7.16) is equivalent to the time-varying ARIMA model

$$y_t - 2y_{t-1} + y_{t-2} = \varepsilon_t + \beta_{1t-1}\varepsilon_{t-1}$$

with β_{1t-1} given by (7.18). After one recursion $P_{2t}=0$ and in the limit $P_{3t}=0$ so that the steady state equivalence is

$$y_t - 2y_{t-1} + y_{t-2} = \varepsilon_t - \varepsilon_{t-1}. \tag{7.19}$$

The z-transfer function of (7.16) is, from (3.10),

$$H(z) = (1\ \ 1)\begin{pmatrix} zI - \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \end{pmatrix}^{-1} z\begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$= \frac{z(z-1)}{(z-1)^2} = \frac{z}{z-1}$$

so that $$Y(z) = \frac{z}{z-1}W(z). \tag{7.20}$$

As Kalman pointed out (Theorem 3.28), only the observable and controllable part of the system can be deduced from (7.20); such a part is

$$\begin{aligned} y_t &= \theta_t \\ \theta_t &= \theta_{t-1} + w_t \end{aligned} \tag{7.21}$$

which has transfer function $z/(z-1)$ as required.

For the model (7.21) $A_t = (C_t+W)/(C_t+W) = 1$, so that $m_t = y_t$ and the forecasts are

$$y_t(k) = y_t \qquad k>1 \tag{7.22}$$

the equivalent ARIMA model being

$$y_t - y_{t-1} = \varepsilon_t. \tag{7.23}$$

However (7.17) is not in general equal to (7.22), since in general $\hat{m} \neq 0$, so that (7.16) is not predictor equivalent to its controllable subsystem. In ARIMA modelling (7.19) would be treated as the model (7.23) with forecasts (7.22) and so similar remarks apply.

7.4 First Order Models

The general univariate DLM has the form

$$y_t = f\theta_t + v_t \tag{7.24}$$

$$\theta_t = g\theta_{t-1} + w_t \tag{7.25}$$

where f and g are scalars. This is observable and controllable provided that f and g are non-zero, and will converge to a steady state solution in the sense of Theorem 4.6 if $W = \mathrm{Var}(w_t) \neq 0$. In this case the equivalent ARIMA model is (cf Theorem 6.14)

$$y_t - gy_{t-1} = \varepsilon_t + \beta\varepsilon_{t-1} \tag{7.26}$$

however the models are restricted in that β can only cover part of its allowable region $|\beta| < 1$. Indeed, from Section 6.6, β has the same sign as $\phi_1 = -g$ and $0 \leq |\beta| \leq |\phi_1|$, so for positive g, $-g<\beta<0$. A much used example of (7.24), (7.25) is the Harrison-Stevens steady model, where f=g=1, which corresponds to the restricted IMA(1,1) model

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

with $-1<\beta<0$, which has been remarked upon by others (eg Godolphin and Stone (1980)).

Before the steady state is reached β in (7.26) depends upon t, as detailed in Theorem 6.23.

We now consider ARIMA(p,d,q) models with q=1, p+d=1 and show that it is possible to construct equivalent DLMs that are unrestricted.

IMA(1,1) A general DLM that is predictor equivalent to an IMA(1,1) must have s=1, $\phi_1=-1$, and to avoid restrictions on $\beta_2 \ldots \beta_{r+s}$ put r=0. Consequently it follows from the discussion of §6.6 that a better choice than the above n=1 non-singular case is n=r+s+1=2, with

$$\underline{G}^2 - \underline{G} = 0. \tag{7.27}$$

The general model is given from (6.69)

$$\begin{aligned} y_t &= (1 \quad 0)\theta(\underline{G}^*)\underline{\theta}_t + v_t \\ \underline{\theta}_t &= \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}\underline{\theta}_{t-1} + \underline{w}_t \end{aligned} \tag{7.28}$$

where $\theta(x)$ is coprime with $x(x-1)$ and so can be taken to be $x+a$, $a\neq 0$, $a\neq -1$; $\underline{G}^*$ is the 2×2 matrix in (7.28). Thus

$$y_t = (1 \quad 0)\begin{pmatrix} a & 1 \\ 0 & 1+a \end{pmatrix}\underline{\theta}_t = (a \quad 1)\underline{\theta}_t \tag{7.29}$$

which with a=1 gives, on putting $\theta_{1t}=\theta_{2t}$, $\theta_{2t}=\theta_{1t}$

$$\begin{aligned} y_t &= (1 \quad 1)\begin{pmatrix} \theta_{1t} \\ \theta_{2t} \end{pmatrix} + v_t \\ \begin{pmatrix} \theta_{1t} \\ \theta_{2t} \end{pmatrix} &= \begin{pmatrix} \theta_{1\,t-1} \\ \theta_{1\,t-1} \end{pmatrix} + \begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix} \end{aligned} \tag{7.30}$$

the model given in Godolphin and Stone (1980), Stone (1982).

For this model $z_t=y_t-y_{t-1}=v_t-v_{t-1}+w_{1t}+w_{1t-1}+w_{2t}-w_{2t-1}$ so that if w_{1t} and w_{2t} are independent with variances W_i

$$\rho_1 = \frac{W_1 - (V+W_2)}{2(V+W_1)+2W_2}. \tag{7.31}$$

The invertibility region is $|\beta|<1$ or $|\rho|<\frac{1}{2}$ and it is

immediate from (7.31) that by altering V, W_1 and W_2 we can cover the full region. The DLM (7.30) is clearly observable and controllable thus assuring us of all the desirable properties, and we have covered the region with a model whose covariance matrix is diagonal.

In fact if we put $W_2=0$ so that $\rho_1 = (W_1-V)/(2V+2W_1)$ then this still covers the region $|\rho|<\frac{1}{2}$ and under the equivalence yields

$$\beta = \frac{W+V \pm \sqrt{(4VW)}}{V-W} \qquad \text{chosen so that } |\beta|<1.$$

By redefining $\theta_{1t}=\theta_t$, $\theta_{2t}=\theta_{t-1}$ the model can be conveniently expressed in the form (with $W_2 \equiv 0$)

$$\begin{aligned} y_t &= \theta_t + \theta_{t-1} + v_t \\ \theta_t &= \theta_{t-1} + w_t \end{aligned} \qquad (7.32)$$

where θ_t can now be thought of as an underlying level - as is the case for the restricted Harrison-Stevens steady model. This expression has the added advantage of being parsimonious with respect to the unknown parameters.

Both (7.31) and (7.32) are much more appealing ways of covering the invertibility region than having v_t and w_t correlated in (7.24), (7.25).

Note that the convex polytope in this case is the extended region $|\rho|\leq\frac{1}{2}$, which the models can also cover.

Example 7.6 - ARMA(1,1)

The Box-Jenkins ARMA(1,1) model is given by

$$y_t - \alpha y_{t-1} = \varepsilon_t + \beta\varepsilon_{t-1}. \qquad (7.33)$$

By a similar argument to above, $\underline{G}^2 = \alpha\underline{G}$ in the case of n=2 so that $\underline{G}^* = \begin{pmatrix} 0 & 1 \\ 0 & \alpha \end{pmatrix}$ and the general model is of the form

$$y_t = (1 \quad 0)\underline{\theta}_t + v_t \tag{7.34}$$

or
$$y_t = (b \quad 1)\underline{\theta}_t + v_t \tag{7.35}$$

with
$$\underline{\theta}_t = \begin{pmatrix} 0 & 1 \\ 0 & \alpha \end{pmatrix}\underline{\theta}_{t-1} + w_t \tag{7.36}$$

because $\Theta(x)$ is of the form $x+b$, $b\neq 0$, $b\neq -\alpha$.

We know from Section 7.2 that a 'canonical form' is $z_t = w^*_{1t} + w^*_{2t} + v_t - \alpha v_{t-1}$, and by using the knowledge gained from the IMA(1,1) example it is clear that putting

$$\begin{aligned} w^*_{1t} &= \hat{w}_{1t} - \hat{w}_{2t} \\ w^*_{2t} &= \hat{w}_{1t} + \hat{w}_{2t} \end{aligned} \tag{7.37}$$

with $\hat{w}_{1t}$, $\hat{w}_{2t}$ independent enables us to cover the full region indeed we can do so without the v_t term which illustrates the remark made at the end of Section 7.2 concerning the 'redundancy' of V. In the notation of that section $\underline{\gamma} = \begin{pmatrix} \underline{F} \\ \underline{F}\,\underline{G} - \alpha\underline{F} \end{pmatrix}$, and it follows that any DLM of the form

$$\begin{aligned} y_t &= (b \quad 1)\underline{\theta}_t + v_t \qquad \{\text{or } (1 \quad 0)\underline{\theta}_t + v_t\} \\ \underline{\theta}_t &= \begin{pmatrix} 0 & 1 \\ 0 & \alpha \end{pmatrix}\underline{\theta}_{t-1} + \underline{\gamma}^{-1}\begin{pmatrix} w^*_{1t} \\ w^*_{2t} \end{pmatrix} \end{aligned} \tag{7.38}$$

will yield the canonical form. For example if $b=1$, $\underline{\gamma} = \begin{pmatrix} 1 & 1 \\ -\alpha & 1 \end{pmatrix}$ and substituting in (7.38) using (7.37) gives

$$y_t = (1 \quad 1)\underline{\theta}_t + v_t \tag{7.39}$$

$$\underline{\theta}_t = \begin{pmatrix} 0 & 1 \\ 0 & \alpha \end{pmatrix}\underline{\theta}_{t-1} + \begin{pmatrix} 0 & -2 \\ 1+\alpha & 1-\alpha \end{pmatrix}\begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix} \tag{7.40}$$

or redefining $\underline{\theta}_t$

$$\underline{\theta}_t = \begin{pmatrix} \alpha & 0 \\ 1 & 0 \end{pmatrix}\underline{\theta}_{t-1} + \begin{pmatrix} 1+\alpha & 1-\alpha \\ 0 & -2 \end{pmatrix}\begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix} \tag{7.41}$$

where w_{1t}, w_{2t} are independent ($w_{it}=w^*_{it}/\{1+\alpha\}$). (7.39) and (7.41) gives the model (7.30) when $\alpha=1$.

Using the equivalence between $\{\underline{F},\underline{G},\underline{R}\,\underline{W}^T\,\underline{R}^T\}$ and $\{\underline{F}\,\underline{R},\,\underline{R}^{-1}\underline{G}\,\underline{R},\underline{W}\}$ gives the equivalent model

$$\begin{aligned} y_t &= (1 \quad 1)\underline{\theta}_t + v_t \\ \underline{\theta}_t &= \tfrac{1}{2}\begin{pmatrix} \alpha+1 & \alpha-1 \\ \alpha+1 & \alpha-1 \end{pmatrix}\underline{\theta}_{t-1} + \begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix}. \end{aligned} \tag{7.42}$$

Alternatively using (7.34) gives by a similar argument the model

$$\begin{aligned} y_t &= (1 \quad 0)\underline{\theta}_t + v_t \\ \underline{\theta}_t &= \begin{pmatrix} 0 & 1 \\ 0 & \alpha \end{pmatrix}\underline{\theta}_{t-1} + \begin{pmatrix} 1 & -1 \\ \alpha+1 & 1-\alpha \end{pmatrix}\begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix}. \end{aligned} \tag{7.43}$$

All the unrestricted models (7.38) - (7.43) differ from the Harrison-Stevens model

$$\begin{aligned} y_t &= \theta_t + v_t \\ \theta_t &= \alpha\theta_{t-1} + w_t \end{aligned} \tag{7.44}$$

for which $z_t=w_t+v_t-\alpha v_{t-1}$, so $\rho_1 = -\alpha V/\{W+V(1+\alpha^2)\}$ yielding the restricted region

$$\frac{-\alpha}{1+\alpha^2} \leq \rho_1 \leq 0.$$

It is possible to consider the extended version of (7.44) by analogy with (7.32), namely

$$\begin{aligned} y_t &= \theta_t + \theta_{t-1} + v_t \\ \theta_t &= \alpha\theta_{t-1} + w_t \end{aligned} \tag{7.45}$$

however this region is also somewhat restricted, for $z_t=w_t+w_{t-1}+v_t-\alpha v_{t-1}$, $\rho_1 = (W-\alpha V)/\{2W+V(1+\alpha^2)\}$ implying

$$\frac{-\alpha}{1+\alpha^2} \leq \rho_1 \leq \tfrac{1}{2}.$$

7.5 Second Order Models

We now look at models whose autoregressive part is of order 2. The general ARMA(2,2) model is

$$y_t - (\lambda_1+\lambda_2)y_{t-1} + \lambda_1\lambda_2 y_{t-2} = \varepsilon_t + \beta_1\varepsilon_{t-1} + \beta_2\varepsilon_{t-2} \qquad (7.46)$$

where $|\lambda_1|<1$, $|\lambda_2|<1$ and the roots of $z^2+\beta_1 z+\beta_2$ are less than 1 in modulus. From Chapter 6 it follows that an equivalent time-invariant DLM must have a system matrix $\underline{G}$ satisfying

$$\underline{G}^{r+1}\{\underline{G}^2 - (\lambda_1+\lambda_2)\underline{G} + \lambda_1\lambda_2\} = 0 \qquad (7.47)$$

where the 'best' choice of n is n=r+s+1=r+3 since s=2 in this case. The smallest possible value of n is 3, in which case the general model is

$$y_t = (1 \quad 0 \quad 0)\,\Theta(G^*)\,\underline{\theta}_t + v_t \qquad (7.48)$$

$$\underline{\theta}_t = \underline{G}^*\,\underline{\theta}_{t-1} + \underline{w}_t \qquad (7.49)$$

where

$$\underline{G}^* = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -\lambda_1\lambda_2 & \lambda_1+\lambda_2 \end{pmatrix}$$

and $\Theta(x)$ is coprime with $x\{x^2-(\lambda_1+\lambda_2)x + \lambda_1\lambda_2\}$ so that the three possibilities are

(i) $\Theta(x) = 1$

(ii) $\Theta(x) = x - a$ $\qquad a\neq 0,\lambda_1,\lambda_2$

(iii) $\Theta(x) = (x - a_1)(x - a_2)$ $\qquad a_i\neq 0,\lambda_1,\lambda_2.$

By redefining the w_t's in (7.6) these models can be written in the 'canonical' form

$$z_t = y_t-(\lambda_1+\lambda_2)y_{t-1}+\lambda_1\lambda_2 y_{t-2} = w_{1t}+w_{2t-1}+w_{3t-2} + \qquad (7.50)$$

$$+ \{v_t-(\lambda_1+\lambda_2)v_{t-1}+\lambda_1\lambda_2 v_{t-2}\}$$

giving the following autocovariances for z_t

$$\gamma_0 = W_{11} + W_{22} + W_{33} + V\{1 + (\lambda_1+\lambda_2)^2 + \lambda_1^2\lambda_2^2\}$$
$$\gamma_1 = W_{12} + W_{23} - (\lambda_1+\lambda_2)(1 + \lambda_1\lambda_2)\, V \qquad (7.51)$$
$$\gamma_2 = W_{13} + \lambda_1\lambda_2$$

with autocorrelations $\rho_k = \gamma_k/\gamma_0$, k=1,2.

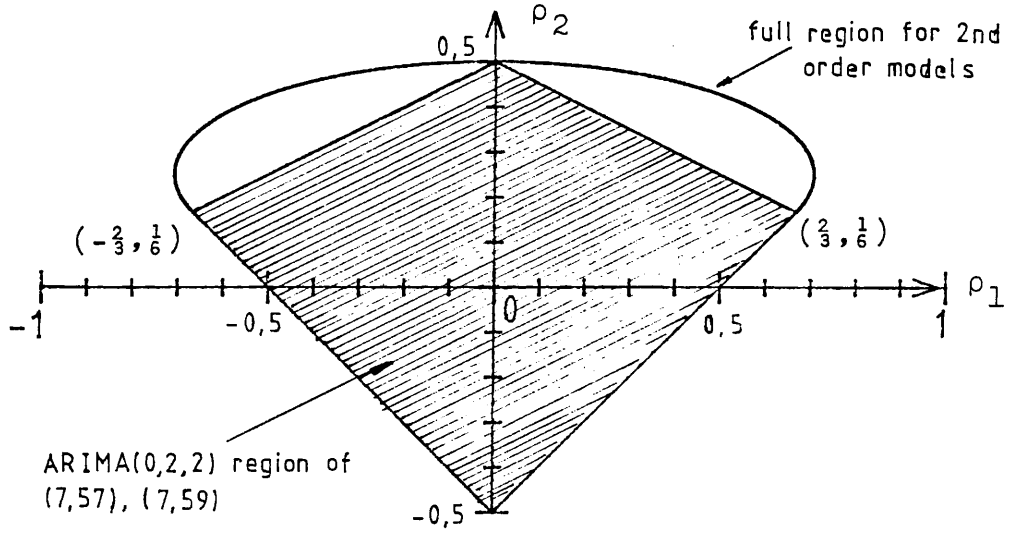
It is well known (for example Box-Jenkins 1970, p 71) that the invertibility (stability) region in the auto-correlation space for the model (7.46) is

$$|\rho_2|<\tfrac{1}{2}, \quad |\rho_1|<\rho_2+\tfrac{1}{2}, \quad |\rho_1|^2<4\rho_2(1-2\rho_2) \text{ for } \rho_2>\tfrac{1}{6} \qquad (7.52)$$

with $\rho_1 = \dfrac{\beta_1(\beta_1+\beta_2)}{1+\beta_1{}^2+\beta_2^2}$, $\rho_2 = \dfrac{\beta_2}{1+\beta_1{}^2+\beta_2{}^2}$.

For convenience we call the region (7.52) together with its limit points the extended invertibility region, so that the inequalities in (7.52) are no longer strict. This corresponds to allowing the roots of $z^2+\beta_1 z+\beta_2$ to have modulus one. Both regions have the appearance of Figure 7.1 where the boundaries are either included or excluded.

FIGURE 7.1: STABILITY REGIONS



The invertibility region can be thought of as the sum of two areas; the first is the triangle bounded by the vertices $(\pm\frac{2}{3}, \frac{1}{6})$, $(0, -\frac{1}{2})$ which is the convex hull of the three points. The second is the remainder bounded by the curve $|\rho_1|^2 < 4\rho_2(1-2\rho_2)$ and the line $\rho_2 = \frac{1}{6}$. It is obvious that the whole region is not a convex polytope, since topologically it has an infinite number of boundary points which do not lie on a finite collection of straight lines. It therefore follows from Theorem 7.4 that it is impossible to cover the full invertibility region with a DLM that has a diagonal covariance matrix, in the sense that the image under R_{II} is not the whole region.

If we put $\underline{W} = \underline{B} \begin{pmatrix} W_1 & 0 & 0 \\ 0 & W_2 & 0 \\ 0 & 0 & W_3 \end{pmatrix} \underline{B}^T$ then from Theorem 7.3 or directly from (7.50), (7.51),

$$\begin{aligned} c_0\rho_1 &= W_1(b_{11}b_{21}+b_{21}b_{31})+W_2(b_{12}b_{22}+b_{22}b_{32}) \quad + \\ &\quad + W_3(b_{13}b_{23}+b_{23}b_{33}) - (\lambda_1+\lambda_2)(1+\lambda_1\lambda_2)V \\ c_0\rho_2 &= W_1 b_{11}b_{31}+W_2 b_{12}b_{32}+W_3 b_{13}b_{33}+\lambda_1\lambda_2 V \end{aligned} \tag{7.53}$$

where $c_0 = W_1(b_{11}^2+b_{21}^2+b_{31}^2)+W_2(b_{12}^2+b_{22}^2+b_{32}^2)+W_3(b_{13}^2+b_{23}^2+b_{33}^2)+$

$$+ V\{1+(\lambda_1+\lambda_2)^2+\lambda_1^2\lambda_2^2\}.$$

Theorem 7.4 tells us that as W_i and V are allowed to vary over the non-negative reals the region traced out is the convex hull of the four points

$$\left(\frac{b_{1i}b_{2i} + b_{2i}b_{3i}}{b_{1i}^2+b_{2i}^2+b_{3i}^2} , \frac{b_{1i}b_{3i}}{b_{1i}^2+b_{2i}^2+b_{3i}^2} \right) \quad i=1,2,3 \tag{7.54}$$

$$\text{and } \left(\frac{-(\lambda_1+\lambda_2)(1+\lambda_1\lambda_2)}{1+(\lambda_1+\lambda_2)+\lambda_1^2\lambda_2^2} , \frac{\lambda_1\lambda_2}{1+(\lambda_1+\lambda_2)+\lambda_1^2\lambda_2^2} \right) \qquad (7.55)$$

which depends upon the λ_i. For the convex hull to cover as large an area as possible the points (7.54) should coincide with boundary points of the full invertibility region. For example putting $b_{1i}=b_{3i}=b_{2i}/\sqrt{2}$ gives the point $(\sqrt{\tfrac{1}{2}}, \tfrac{1}{4})$; (7.54) gives $(0, \pm\tfrac{1}{2})$ if and only if $b_{2i}=0$, $b_{1i}=\pm b_{3i}$ and the value $(\pm\tfrac{2}{3}, \tfrac{1}{6})$ if $b_{1i}=b_{3i}$, $b_{2i}=\pm 2b_{1i}$. Thus if

$$\underline{B} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 2 \\ 1 & -1 & 1 \end{pmatrix} \qquad (7.56)$$

then the three points (7.54) are $(0,\tfrac{1}{2})$, $(0,-\tfrac{1}{2})$, $(\tfrac{2}{3}, \tfrac{1}{6})$. The region depends upon λ_i because (7.55) does. In the case $\lambda_1=\lambda_2=1$, giving the ARIMA(0,2,2) model, then (7.55) is $(-\tfrac{2}{3}, \tfrac{1}{6})$ and the region is the shaded area in Figure 7.1. This is the best that can be done in this case.

For different values of λ_i the areas covered shrink as the point (7.55) moves. The smallest region covered in such a case is the triangle bounded by the three points $(0,\pm\tfrac{1}{2})$, $(\tfrac{2}{3}, \tfrac{1}{6})$.

The models mentioned above can be realised by applying the procedure of Example 7.6, for instance

$$y_t = (1 \quad 0 \quad 0)\underline{\theta}_t + v_t \qquad (7.57)$$

$$\underline{\theta}_t = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -\lambda_1\lambda_2 & \lambda_1+\lambda_2 \end{pmatrix} \underline{\theta}_{t-1} + \underline{\gamma}^{-1} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 2 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} w_{1t} \\ w_{2t} \\ w_{3t} \end{pmatrix}$$

where $\underline{\gamma}$ is obtained from (6.60) as

$$\underline{\gamma} = \begin{pmatrix} \underline{F} \\ \underline{F}\,\underline{G} - (\lambda_1+\lambda_2)\underline{F} \\ \underline{F}\,\underline{G}^2 - (\lambda_1+\lambda_2)\underline{F}\,\underline{G} + \lambda_1\lambda_2\underline{F} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -(\lambda_1+\lambda_2) & 1 & 0 \\ \lambda_1\lambda_2 & -(\lambda_1+\lambda_2) & 1 \end{pmatrix} \tag{7.58}$$

Substituting $\lambda_i=1$ in (7.57), (7.58) gives the evolution

$$\underline{\theta}_t = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 2 \end{pmatrix} \underline{\theta}_{t-1} + \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 4 \\ 4 & 2 & 8 \end{pmatrix} \begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix} \tag{7.59}$$

where the w_{it} are independent. This can be used to give an equivalent DLM that has a diagonal covariance matrix by proceeding as in Example 7.6.

The region covered depends upon λ_1, λ_2. If we wish to cover the shaded area in Figure 7.1 for all λ_i then the dimension of the system vector must be increased - to at least 4. For n=4 the matrix $\underline{G}$ satisfies (7.47) with r=1, but for this to be equivalent to an ARIMA model with moving average part of order 3 we require additionally $\beta_3=0$ (or $\rho_3=0$) which places restrictions on the matrix $\underline{B}$. The exact form can be obtained by a similar procedure to the above. Only by having a system vector of size n with n large can we start to cover all of the invertibility region (unless we allow $\underline{W}$ to range over the positive semi-definite matrices).

Example 7.7

The Harrison-Stevens ARMA(2,2) model is

$$\begin{aligned} y_t &= (1 \quad 0)\underline{\theta}_t + v_t \\ \theta_{1t} &= \lambda_1\theta_{1t-1} + \theta_{2t} + w_{1t} \\ \theta_{2t} &= \lambda_2\theta_{2t-1} + w_{2t} \end{aligned} \tag{7.60}$$

This has a non-singular system matrix when expressed as a DLM so the invertibility region is restricted.

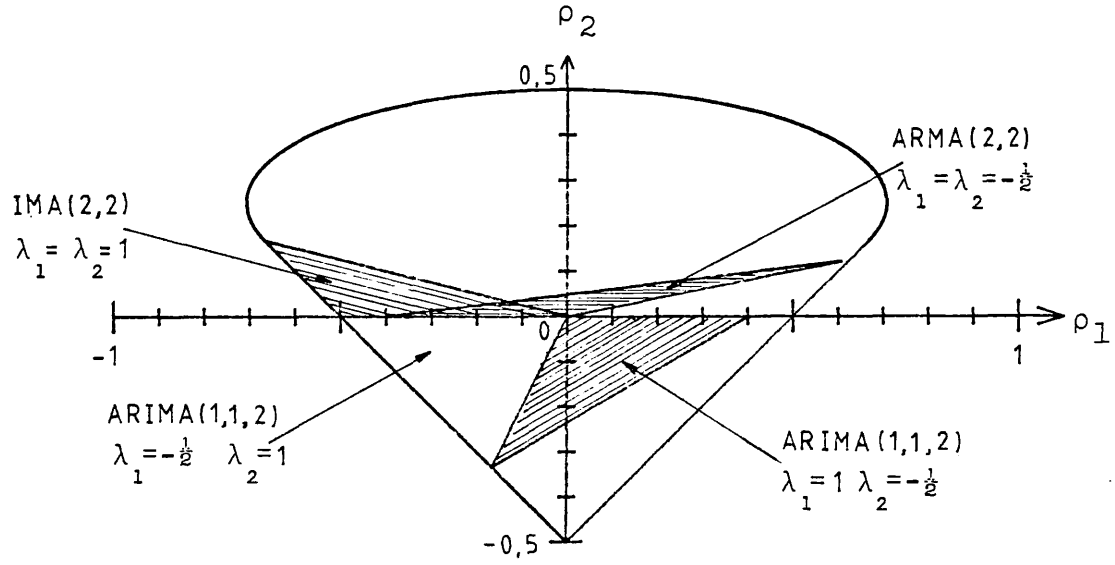
It follows directly from (7.60) that

$$z_t \triangleq y_t-(\lambda_1+\lambda_2)y_{t-1}+\lambda_1\lambda_2 y_{t-1} = w_{1t}-\lambda_2 w_{1t-1}+w_{2t}+v_t -(\lambda_1+\lambda_2)v_{t-1}+\lambda_1\lambda_2 v_{t-2}$$

$\rho_1 = \{-\lambda_2 W_1 - (\lambda_1+\lambda_2)(1+\lambda_1\lambda_2)V\}/D$, $\rho_2 = \lambda_1\lambda_2 V/D$, where $D = W_1(1+\lambda_2^2) + W_2 + V\{1+(\lambda_1+\lambda_2)^2+\lambda_1^2\lambda_2^2\}$. Theorem 7.4 says that the image under R_{II} is the convex hull of the vertices $\left(-\frac{\lambda_2}{1+\lambda_2^2}, 0\right)$, $(0, 0)$ and $\left\{-\frac{(\lambda_1+\lambda_2)(1+\lambda_1\lambda_2)}{D^*}, \frac{\lambda_1\lambda_2}{D^*}\right\}$ with $D^* = \{1+(\lambda_1+\lambda_2)^2+\lambda_1^2\lambda_2^2\}$. These points are obviously heavily dependent upon the λ_i and the region is much more severely restricted than the preceding models in this section. Some of the regions are sketched in Figure 7.2.

FIGURE 7.2: INVERTIBILITY REGIONS FOR EXAMPLE 7.7

Harrison-Stevens models



It is possible to proceed in an exactly analagous way to this and the last section and consider higher order ARIMA models, although the invertibility regions become more complex.

We have derived 'canonical forms' in the above by confining attention to diagonal covariance matrices. Note however that any fixed ARMA/model (i.e. fixed parameters) can be represented by a DLM with a fixed diagonal covariance matrix. This is because we know from Theorem 7.2 that we can cover the invertibility region with a DLM that has a non-negative covariance matrix, $\{\underline{F}, \underline{G}, \underline{W}\}$ say. But then $\underline{W}$ is diagonisable as $\underline{R}\,\underline{W}\,\underline{R}^T = \text{diag}\{W_i\}$ for some non-singular $\underline{R}$ and where W_i are specific values, and so the strictly equivalent DLM $\{\underline{F}\,\underline{R}^{-1}, \underline{R}\,\underline{G}\,\underline{R}^{-1}, \underline{R}\,\underline{W}\,\underline{R}^T\}$ proves the assertion.

7.6 Augmented DLM's and Summary

Example 7.7 and (7.44) are two instances of a much broader class of DLMs used by Harrison and Stevens (1971, 1976). These are the so-called extended Markov polynomial models, which are given by

$$y_t = \theta_{1t} + v_t \tag{7.61}$$

$$\begin{aligned} \theta_{1t} &= \lambda_1\theta_{1t-1} + \theta_{2t} + w_{1t} \\ \theta_{2t} &= \lambda_2\theta_{2t-1} + \theta_{3t} + w_{2t} \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots \\ \theta_{nt} &= \lambda_n\theta_{nt-1} + w_{nt} \end{aligned} \tag{7.62}$$

which can be expressed as the DLM

$$y_t = (1 \quad 0 \quad \ldots \quad 0)\underline{\theta}_t + v_t \tag{7.63}$$

$$\underline{\theta}_t = \begin{pmatrix} \lambda_1 & \lambda_2 & \ldots & \lambda_n \\ 0 & \lambda_2 & & \\ & & \ddots & \\ 0 & 0 & & \lambda_n \end{pmatrix} \underline{\theta}_{t-1} + \begin{pmatrix} 1 & 1 & \ldots & 1 \\ & 1 & \ldots & 1 \\ \underline{0} & & \ddots & \\ & & & 1 \end{pmatrix} \begin{pmatrix} w_{1t} \\ w_{2t} \\ \vdots \\ w_{nt} \end{pmatrix} \tag{7.64}$$

where the w_i's are independent and the λ_i non-zero, thus this is an example of the preferred covariance structure.

For this class of models the observability matrix is $\underline{M} = \{m_{ij}\}$ with $m_{11}=1$, $m_{1j}=0$ for $j \neq 1$, and $m_{ij} = \lambda_j(m_{i-1\,1} + m_{i-1\,2} + \ldots + m_{i-1\,j})$ otherwise. Expanding by the first row gives the determinant

$$|\underline{M}| = \begin{vmatrix} m_{22} & \ldots & m_{2n} \\ \vdots & & \\ m_{m2} & \ldots & m_{nn} \end{vmatrix} = \lambda_2\lambda_3\ldots\lambda_n \begin{vmatrix} 1 & 1 & \ldots & 1 \\ b_{32} & b_{33} & \ldots & b_{3n} \\ \vdots & & & \\ b_{n2} & b_{n3} & \ldots & b_{nn} \end{vmatrix}$$

where $b_{ij} = m_{ij} / \lambda_j$. Subtracting the j^{th} column from the $j+1^{th}$, for $j=n-1, \ldots 1$, with the relation $b_{ij+1} - b_{ij} = m_{i-1\,j+1}$ and expanding by the first row gives

$$|\underline{M}| = \lambda_2\lambda_3 \ldots \lambda_n \begin{vmatrix} m_{23} & \ldots & m_{2n} \\ \vdots & & \\ m_{n-13} & \ldots & m_{n-1n} \end{vmatrix}$$

But this is $\lambda_2\ldots\lambda_n$ multiplying the determinant of an n-1 order model with parameters $\lambda_2, \ldots \lambda_n$, and so a simple inductive argument proves that $|M| = \lambda_2\lambda_3^2\ldots\lambda_n^{n-1}$, hence

Theorem 7.8

The system (7.61),(7.62) or (7.63),(7.64) is observable if and only if $\lambda_i \neq 0$ for $i \geq 2$, and has a non-singular system matrix if in addition $\lambda_1 \neq 0$.

For this latter class of models the characteristic equation is $\prod_{1}^{n} (\lambda - \lambda_i)$ which is also the minimal polynomial. The model is controllable and in the steady state is equivalent to

$$\prod_{1}^{n} (1 - \lambda_i B) y_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_n \varepsilon_{t-1} ,$$

an ARIMA(n-d,d,n) model where d is the number of λ_i's equal to one. Since the system matrix is non-singular these models can only describe very restricted regions in the correlation or parameter spaces.

However such models do have the advantage of easy interpretability, in contrast to those of the previous sections where the emphasis was on finding models which possess desirable mathematical properties. For example if all the λ_i are equal to one, then in (7.61) θ_{1t} can be thought of as an underlying level which evolves in a Markov fashion, namely is a noisy version of the previous level augmented by a slope term, θ_{2t}. The slope is the first difference of the level - a discrete version of the derivative- and similarly for the other terms in (7.62).

Ideally we would like to have DLM's that have an easy interpretation and which cover the entire stability region of the equivalent ARIMA model. This is the case for IMA(1,1) models as illustrated by (7.32). Motivated by this we look at models

$$y_t = \theta_{1t} + \theta_{1t-1} + v_t \qquad (7.65)$$

where θ_{1t} evolves as (7.62) - or (7.64). The model can be witten in DLM form by defining $\theta_{n+1\,t} = \theta_{1\,t-1}$, with matrices $\underline{F}^*$, $\underline{G}^*$ and $\underline{B}^*$ where

$$\underline{F}^* = (\underline{F} \quad 0), \quad \underline{G}^* = \begin{pmatrix} \underline{G} & 0 \\ 1\ 0\ \ldots\ 0 & 0 \end{pmatrix}, \quad \underline{B}^* = \begin{pmatrix} \underline{B} \\ 0\ \ldots\ 0 \end{pmatrix} \tag{7.66}$$

and where $\underline{F}$, $\underline{G}$ and $\underline{B}$ are the matrices in (7.64). The observability matrix for (7.66) $\underline{M}^*$ has determinant

$$|\underline{M}^*| = \begin{vmatrix} \underline{F} & 1 \\ \underline{FG} + \underline{F} & 0 \\ \vdots & \vdots \\ \underline{FG}^n + \underline{FG}^{n-1} & 0 \end{vmatrix} = \begin{vmatrix} \underline{F} \\ \underline{FG} \\ \vdots \\ \underline{FG}^{n-1} \end{vmatrix} |\underline{G} + \underline{I}| = |\underline{M}|\ |\underline{G} + \underline{I}|.$$

Using this result with Theorem 7.8 proves

Theorem 7.9

The augmented model (7.66) is observable provided that $\underline{G} + \underline{I}$ has full rank, where $\underline{G}$ is the system matrix in (7.64), that is if and only if $\lambda_i \neq -1$ for any i. Moreover under these conditions the model is controllable.

The proof of controllability follows from the fact that $(\underline{B}^* \quad \underline{G}^*\underline{B}^*)$ has rank n+1 since the first column of $\underline{G}^*\underline{B}^*$ is $(\lambda_1\ 0\ \ldots\ 0\ 1)$ which is linearly independent of the n columns of $\underline{B}^*$.

Thus we have a class of 'augmented' Markov polynomial models which are observable and controllable, which have a diagonal covariance matrix and which cover more of the invertibility region than the unaugmented versions.

Example 7.10

An augmented version for the ARMA(2,2) model is

$$\begin{aligned} y_t &= \theta_{1t} + \theta_{1t-1} + v_t \\ \theta_{1t} &= \lambda_1\theta_{1t-1} + \theta_{2t} + w_{1t} \end{aligned} \tag{7.67}$$

$$\theta_{2t} = \lambda_2\theta_{2t-1} + w_{2t} \qquad (7.68)$$

thus the derived process

$$z_t = w_{1t}+(1-\lambda_2)w_{1t-1}-\lambda_2 w_{1t-1}+w_{2t}+w_{2t-1}+v_t-(\lambda_1+\lambda_2)v_{t-1}+\lambda_1\lambda_2 v_{t-2}$$

has autocorrelations

$$\rho_1 = \{W_1(1-\lambda_2)^2 + W_2 - V(\lambda_1+\lambda_2)(1+\lambda_1\lambda_2)\} / D \qquad (7.69)$$

$$\rho_2 = (-\lambda_2 W_1 + \lambda_1\lambda_2 V) / D \qquad (7.70)$$

where $D = W_1\{1+(1-\lambda_2)^2+\lambda_2^2\} + 2W_2 + V\{1+(\lambda_1+\lambda_2)^2+\lambda_1^2\lambda_2^2\}$

and the vertices of the convex hull are therefore

$$(\tfrac{1}{2}, 0),\ \left(\frac{(1-\lambda_2)^2}{D_1}, \frac{-\lambda_2}{D_1}\right),\ \left(\frac{-(\lambda_1+\lambda_2)(1+\lambda_1\lambda_2)}{D_v}, \frac{\lambda_1\lambda_2}{D_v}\right) \qquad (7.71)$$

with $D_1 = 1+(1-\lambda_2)^2+\lambda_2^2$, $D_v = 1+(\lambda_1+\lambda_2)^2+\lambda_1^2\lambda_2^2$.

Some of the appropriate regions in the autocorrelation space for various values of the λ_i are sketched in Figures 7.3 to 7.6, using (7.71).

Box-Jenkins place great emphasis on parsimony, which in practice means that the class of ARIMA(p,d,q) models considered is a very small one -for which p, d and q are usually less than 2 or 3. The results of this Section and the previous ones as they apply to most of these models are summarised in Table 7.1.

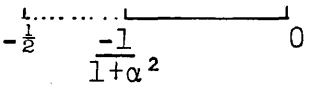
In Table 7.1 the ARIMA(1,1,2) and (0,2,2) models can be obtained from those for the ARMA(2,2) model by putting $\lambda_2=1$ or $\lambda_1=\lambda_2=1$ respectively. We allow for some of the β's to be zero to include ARIMA(p,d,q) models with $q<p+d$. (7.74), (7.75) illustrate the general result that Markov polynomial models and their augmented versions depend upon

the particular permutation of $\lambda_1 \ldots \lambda_n$ used whereas the ARIMA models are independent of such ordering.

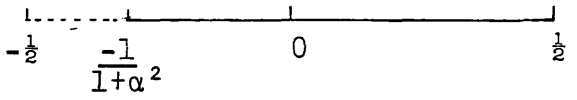
The invertibility regions for these models (that is for the derived processes z_t in the correlation plane) are

1. ARMA(1,1)

   Harrison-Stevens



   Augmented



2. IMA(1,1) - as above with $\alpha=1$.

3. ARMA(2,2), ARIMA(1,1,2), ARIMA(0,2,2)

The Harrison-Stevens' regions are detailed in Example 7.7 and illustrated in Figure 7.2. The augmented model is given in Example 7.10 and some of the regions are given below in Figures 7.3 - 7.6.

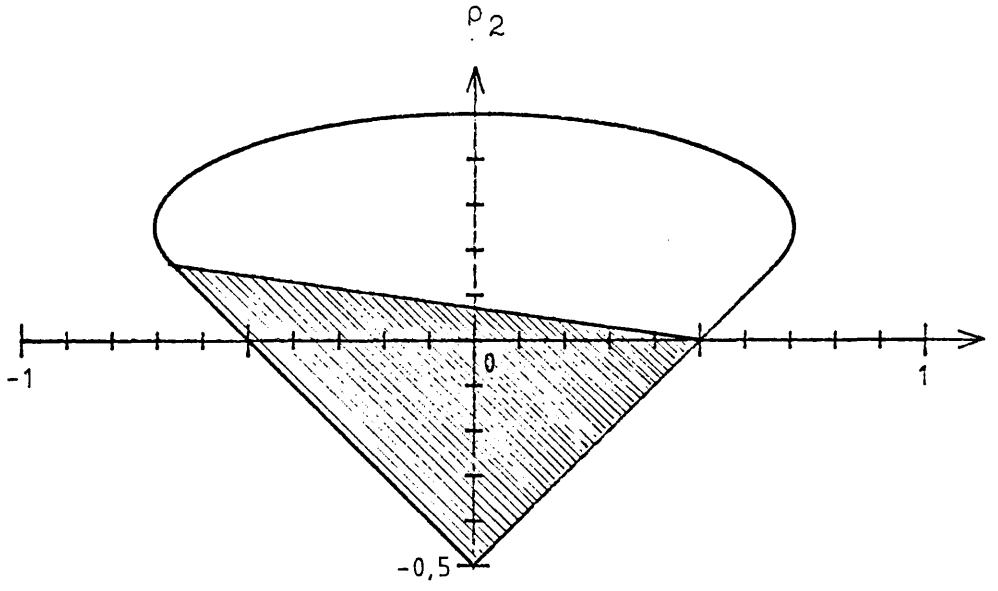


FIGURE 7.3: STABILITY REGION FOR ARIMA(0,2,2) MODEL OF EXAMPLE 7.10

FIGURE 7.4

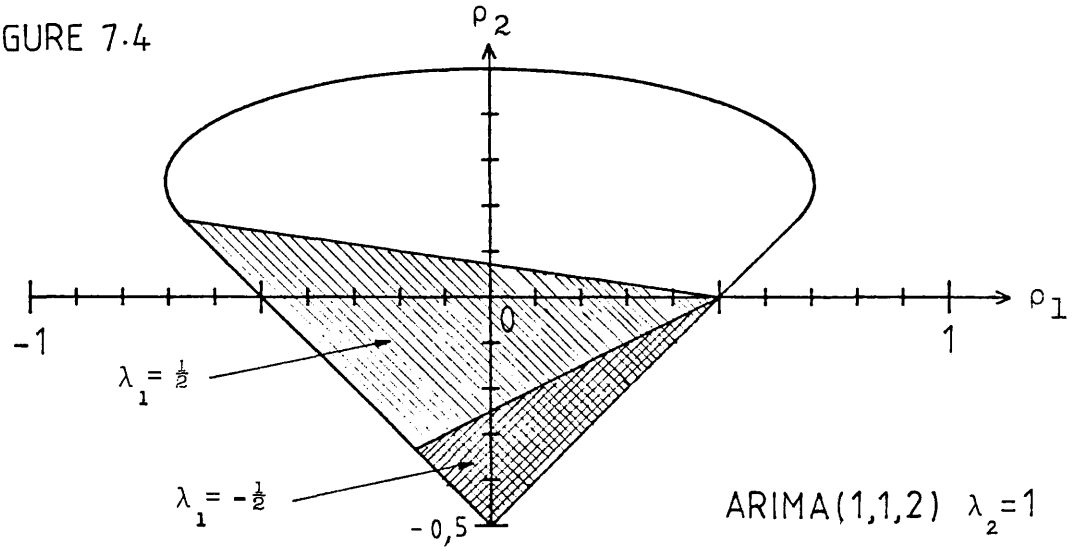


FIGURE 7.5

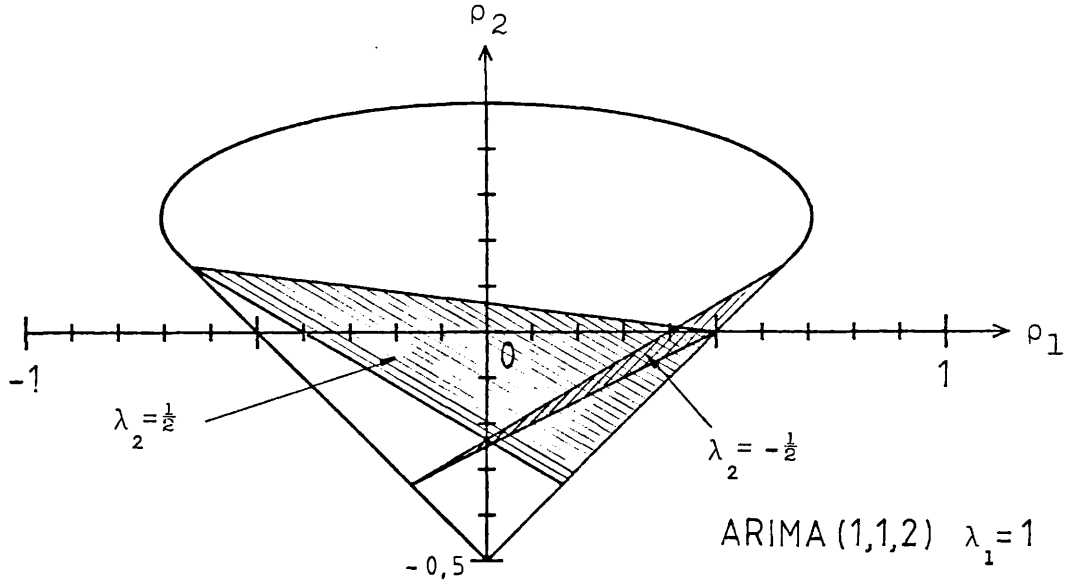
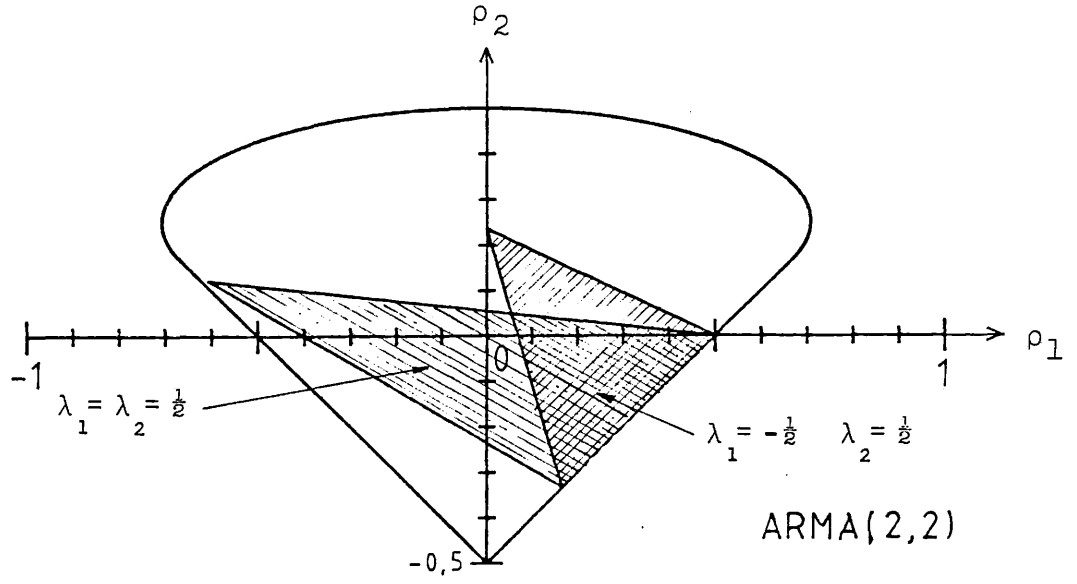


FIGURE 7.6



FIGURES 7.4-7.6: STABILITY REGIONS FOR EXAMPLE 7.10

TABLE 7.1: SUMMARY OF FIRST AND SECOND ORDER MODELS.

1. ARMA(1,1) $y_t - \alpha y_{t-1} = \varepsilon_t + \beta\varepsilon_{t-1}$

Harrison-Stevens
$$y_t = \mu_t + v_t \qquad (7.72)$$
$$\mu_t = \alpha\mu_{t-1} + w_t$$

Augmented model
$$y_t = \mu_t + \mu_{t-1} + v_t \qquad (7.73)$$
$$\mu_t = \alpha\mu_{t-1} + w_t$$

'Full' model
$$y_t = \theta_{1t} + \theta_{2t} + v_t$$
$$\theta_{1t} = \tfrac{1}{2}(\alpha+1)\theta_{1t-1} + \tfrac{1}{2}(\alpha-1)\theta_{2t-1} + w_{1t}$$
$$\theta_{2t} = \tfrac{1}{2}(\alpha+1)\theta_{1t-1} + \tfrac{1}{2}(\alpha-1)\theta_{2t-1} + w_{2t}$$
w_{1t}, w_{2t} independent

2. IMA(1,1) $y_t - y_{t-1} = \varepsilon_t + \beta\varepsilon_{t-1}$

Harrison-Stevens
$$y_t = \mu_t + v_t \qquad (7.74)$$
$$\mu_t = \mu_{t-1} + w_t$$

Augmented / Full
$$y_t = \mu_t + \mu_{t-1} + v_t \qquad (7.75)$$
$$\mu_t = \mu_{t-1} + w_t$$

3. ARMA(2,2) $y_t-(\lambda_1+\lambda_2)y_{t-1}+\lambda_1\lambda_2 y_{t-2} = \varepsilon_t+\beta_1\varepsilon_{t-1}+\beta_2\varepsilon_{t-2}$

Harrison-Stevens
$$y_t = \mu_t + v_t$$
$$\mu_t = \lambda_1\mu_{t-1} + \beta_t + w_{1t} \qquad (7.76)$$
$$\beta_t = \lambda_2\beta_{t-1} + w_{2t}$$

Augmented
$$y_t = \mu_t + \mu_{t-1} + v_t$$
$$\mu_t = \lambda_1\mu_{t-1} + \beta_t + w_{1t} \qquad (7.77)$$
$$\beta_t = \lambda_2\beta_{t-1} + w_{2t}$$

Full
$$y_t = \mu_t + v_t$$
$$\mu_t = \beta_{t-1} + \bar{w}_{1t}$$
$$\beta_t = \gamma_{t-1} + \bar{w}_{2t}$$
$$\gamma_t = (\lambda_1+\lambda_2)\gamma_{t-1} - \lambda_1\lambda_2\beta_{t-1} + \bar{w}_{3t}$$

$\bar{w}_{it}$ not independent

A 'best' model with w_{it} independent is provided in Section 7.6.

7.7 Inference For Augmented DLMs

It is possible to use the standard Kalman Filter on the augmented models provided that we express them in DLM form, as in (7.66) say. However these relations can be slightly simplified by effectively reducing the dimension of the system vector so that it has the same dimension as before augmentation. This is becaues if we write $\underline{C}^*_t = \text{Cov}[\underline{\theta}_t | y^t] = (c_{ij})$, $1 \leq i,j \leq n+1$, the posterior covariance of the augmented system, then in the notation of (7.66) $\underline{P}^*_t = \underline{G}^*\underline{C}^*_t\underline{G}^{*T} + \underline{B}^*\underline{W}^*\underline{B}^{*T}$ is a function of the unaugmented matrices $\underline{G}$, $\underline{B}$, $\underline{W}$ and $\underline{C}_t = (c_{ij})$, $1 \leq j \leq n$. But the Kalman filter depends only on $\underline{P}^*_t$ and $\underline{F}^*\underline{G}^*\underline{m}_t$ which is a function of $m_{1t}, \ldots, m_{nt}$ and so confirms the conjecture. The following example makes the point clear

Example 7.11

The steady augmented model (7.75)

$$y_t = \theta_t + \theta_{t-1} + v_t$$
$$\theta_t = \theta_{t-1} + w_t$$

can be expressed via (7.66) as a DLM with matrices

$$\underline{F} = (1 \quad 1), \quad \underline{G} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad \underline{W} = \begin{pmatrix} W & 0 \\ 0 & 0 \end{pmatrix}$$

and system vector $(\theta_t \quad \theta_{t-1})^T$. From the filter

$$\underline{P}_t = \underline{G}\,\underline{C}_{t-1}\,\underline{G}^T + \underline{W} = \begin{pmatrix} C_{1t-1} + W & C_{1t-1} \\ C_{1t-1} & C_{1t-1} \end{pmatrix}$$

where $C_{1t-1} = (\underline{C}_{t-1})_{11}$. Thus the updates are

$$\hat{y} = 2m_{1t-1}, \qquad \hat{Y} = 4C_{1t-1} + W + V$$

$$\underline{m}_t = \begin{pmatrix} m_{1t} \\ m_{2t} \end{pmatrix} = \begin{pmatrix} m_{1t-1} \\ m_{1t-1} \end{pmatrix} + \begin{pmatrix} 2C_{1t-1}+W \\ 2C_{1t-1} \end{pmatrix} \cdot \frac{(y_t - \hat{y})}{\hat{Y}}$$

and

$$\underline{C}_t = \underline{P}_t - \begin{pmatrix} 2C_{1t-1}+W \\ 2C_{1t-1} \end{pmatrix} (2C_{1t-1}+W \quad 2C_{1t-1}) / \hat{Y}$$

with predictors

$$E[y_{t+k}|y^t] = \underline{F}\,\underline{G}^k \underline{m}_t = 2m_{1t} \tag{7.78}$$

and $\mathrm{Var}(y_{t+k}\, y^t) = 4C_{1t-1} + 4(k-1)W + W + V$.

Therefore provided that we are not interested in the smoothed estimate $\theta_{t-1}|y^t$, the forecasting equations can be simplified to

$$\begin{aligned} \hat{y}_t &= 2m_{t-1} \\ \hat{Y}_t &= 4C_{t-1} + W + V \\ A_t &= \frac{2C_{t-1}+W}{4C_{t-1}+W+V}, \qquad P_t = C_{t-1} + W \\ m_t &= m_{t-1} + A_t(y_t - \hat{y}_t) \\ C_t &= (I-A_t)P_t - A_tC_{t-1} = P_t - A_t(P_t + C_{t-1}) \end{aligned} \tag{7.79}$$

which only involves the two terms m_t and C_t, corresponding to the old m_{1t} and C_{1t}, with the predictors given by (7.78) with these assignments. Although (7.79) is very similar to the Kalman filter updates it is slighlty different, indeed it is not the filter for any scalar DLM.

If the error terms are allowed to take the more general form

$$v_t \sim N(m_v, V) \quad \text{and} \quad w_t \sim N(m_w, W)$$

then the updates are as in (7.79) except for

$$\begin{aligned} \hat{y}_t &= 2m_{t-1} + m_w + m_v \\ m_t &= m_{t-1} + m_w + A_t(y_t - \hat{y}_t). \end{aligned} \tag{7.80}$$

It is possible to interact with state-space models in a simple way, as Harrison and Stevens (1976) mention. For example for the steady model (7.75) we can alter some of the values of m_v, m_w, V or W at some specific time T. For the model, altering m_v or V just affects the current observation, whilst changing m_w alters the underlying level of the process which affects all subsequent time-points. Increasing W allows the level to change more over time, thereby increasing our uncertainty as to its value. If m_w is non-zero for successive time periods then a deterministic growth is introduced into the model.

The implications on the estimators of altering these quantities are a little more hidden. For example a non-zero m_v affects the posterior estimate of the level (and therefore successive estimates as well) through (7.80). A non-zero value of m_w increases the posterior of the level not by m_w but by $m_w(1-A_t)$ which is less than m_w.

Increasing W increases A_t (in (7.79)) so that more weight is given to the current observation in updating the level through

$$m_t = m_{t-1} + A_t(y_t - \hat{y}_t).$$

Conversely increasing V means that the observation contains less information, A_t decreases and less weight is given to the observation.

CHAPTER 8

EFFICIENT COLLAPSING PROCEDURES FOR CLASS II MODELS

8.1 INTRODUCTION

In Chapter 4 Section 7 we showed how given m past histories $H_1, \ldots H_m$ at time t = 1 which lead to a posterior $\underline{\theta}_{t-1}|y^{t-1}$ that is a mixture of m (multivariate) normal distributions, the introduction of an m x n transition matrix $\{\Pi_{ij}(t)\}$ leads to a posterior at time t, $\underline{\theta}_t|y^t$, which is a mixture of mn normal distributions. This obviously creates an explosive situation, for example if we were to introduce n new models at each time stage, then after time T we would have a posterior of n^T components, assuming that we had started from a single component at time zero. Consequently the model becomes prohibitively cumbersome, with more and more computations required, each component having to be updated via the Kalman Filter. In this chapter we comment briefly on the Harrison-Stevens Class II approach to this problem, as outlined in Chapter 4, and then propose a different solution.

8.2 Problems with Class II Models

If $\{\Pi_{ij}(t)\} = \{\Pi_{ij}\}$ is an n x n matrix independent of t, introducing n models each time, then the Harrison-Stevens collapse procedure reduces the number of models entertained from n^2 to n each time so that the posterior is an n component normal mixture. If we denote the history arising from model j at time t and history i at time t - 1

by $M_{ij}(t)$, then as pointed out by Gathercole and Smith (1984), the Harrison-Stevens approach can be thought of as a collapse of the n^2 histories $\{M_{ij}(t)\}$ into n groups $\{G_k\}$ under the relation

$$G_j = \{M_{rs} \text{ s.t. } D_{HS}(M_{ij},M_{rs})=0\}$$

where $D_{HS}(M_{ij},M_{rs})$ is the binary distance measure

$$D_{HS}(M_{ij},M_{rs}) = 0 \quad \text{if} \quad j = s$$
$$= 1 \qquad \text{otherwise.}$$

Each of the n groups G_j is then represented by a single normal density together with an associated probability as detailed in Chapter 4. Such modelling assumes that the underlying system vector has the same dimension for each of the n^2 elements in the precollapsed mixture, if necessary this can be achieved by embedding the models using the largest system vector. If the dimension of this vector is k say, then each of the n^2 elements in the mixture corresponds to a surface in k-dimensional space, or alternatively to n^2 points in $k(k+3)/2+1$ dimensional Euclidean space. $k(k+3)/2$ being the number of free parameters in the normal distribution, the 1 being for its associated probability. Seen in this light, the collapsing procedure is a clustering algorithm which forms n clusters from the n^2 points or surfaces, the cluster centre being chosen to preserve the information within the group. As remarked in Gathercole and Smith (1984), using the method of moments as Harrison-Stevens do to chose the cluster centre means that the increase in uncertainty incurred in representing a cluster by its centre is represented by an increased variance.

The weakness of the above collapsing procedure is that the clusters are determined solely on the labelling of the n^2 histories or points, and takes no account of how 'close' or 'far apart' the points are under sensibly defined distances. Consequently it is not difficult to construct examples where such a collapse would not be sensible, for instance Gathercole and Smith construct an example where multimodal posteriors are approximated by unimodal densities so that the approximation is not even topologically equivalent to the uncollapsed mixture. These authors propose a decision theoretic solution to such problems. The disadvantages of such a method, apart from the fact that the clustering is not necessarily defined in a natural way, is that it depends on the appropriate loss function. Although ultimately any decisions in a Bayesian framework depend on the decision-makers loss function, the Gathercole-Smith approach actually alters the form of the predictive distribution, so that using different loss functions would require complete recalculation. We now outline a whole class of collapsing procedures which are not subject to such restrictions and which have a readier interpretation in terms of a clustering procedure.

The procedure given has links with that of Sorenson and Alspach (1971), where a justification for using Gaussian sums is given, and a simple method of collapsing is given.

8.3 A Class of Collapsing Procedures

Consider the simpler case examined by Harrison and Stevens where $\Pi_{ij}(t)$ is independent of i, that is the

probability that model j is in operation at time t is independent of the past history. At time t we have a posterior of the form

$$\sum_{k=1}^{N} p_k N(\underline{\mu}_k, \underline{C}_k) \text{ where } \sum p_k = 1 \text{ for some integer N and we}$$

wish to approximate this by a normal mixture with a smaller number of components, say

$$\sum_{k=1}^{N^*} p_k^* N(\underline{\mu}_k^*, \underline{C}_k^*) \text{ with } \sum p_k^* = 1 \text{ and } N^* < N.$$

To achieve this we adopt a clustering procedure by using a criterion to form N^* clusters and then defining cluster centres to summarise the information contained in the clusters. Two ways in which we can do this are

Cluster Method 1

If we define $\underline{f}_k = p_k N(\underline{\mu}_k, \underline{C}_k)$ where each $N(\underline{\mu}_k, \underline{C}_k)$ is an r dimensional Normal distribution then $\underline{f}_k$ is an L integrable function with $\int \underline{f}_k = p_k$.

Let $d(\underline{f}, \underline{g})$ to be a metric on the space of L^r integrable functions, and let C be a clustering mapping which is complete in that it is a mapping from i,j onto 0,1 such that

$$c(d_{ij}) = \begin{cases} 0 \text{ if i,j in the same cluster} \\ 1 \text{ otherwise .} \end{cases}$$

Then we can define a clustering procedure c(d) to induce clusters or groups G_k,

$$G_k^* = \{f_k, f_m \mid c(d(f_k, f_m)) = 0\}.$$

It is then a simple matter to define cluster centres by analogy with Harrison and Stevens.

Cluster Method II

Alternatively we can treat the associated probabilities separately, so that we perform a clustering procedure on the probability densities $f_k = N(\underline{\mu}_k, \underline{C}_k)$ enabling us to use a metric d on the space of probability densities together with a clustering prodedure c(d). To avoid too many comparisons we can discard those models whose associated probability falls below a critical value, or as in Gathercole and Smith (1984) omit those whose odds ratio with respect to the most probable model falls below some value; in both cases distributing the probability among the remaining models.

In effect we then have four steps

1 Discard those models whose odds ratio with respect to the most probable falls below a specified critical value and adjust the probabilities accordingly

2 Calculate a distance matrix using a probability matrix d

3 Perform a clustering procedure to produce m clusters

4 Calculate the cluster centres.

It is important that the clustering mapping is complete in that it produces a number of clusters; for example agglomerative clustering produces chains of clusters, so that our clustering C might consist of an agglomerative procedure together with a stopping rule, which produces a definite number of clusters.

It can be seen that provided we define suitable metrics such procedures can be applied to general mixtures - not just Normal mixtures.

We shall use the second cluster method, which has certain advantages in that we can use more tractable metrics, and also mixtures such as $p_1 f + p_2 f$ would be clustered, since (f,f) has distance zero, which is what we want. Conversely method I would define in general non-zero distance between $p_1 f$ and $p_2 f$ and so not necessarily cluster them together. The disadvantage of method II is that it is independent of p_1 and p_2, and there might be cases where $p_1 f$ and $p_2 g$ are close whilst f and g are not.

The clustering method used in Chapter 9 is agglomerative clustering based upon furthest neighbour distance. That is we start from single objects, find the two nearest ones and make them into a cluster. The distances of all objects from this cluster are calculated and the process is repeated by joining the two nearest objects (clusters) until all the clusters are greater than some specified distance apart. The clustering then stops. 'Furthest neighbour' distance defines the distance between two clusters I, J as

$$D = \max_{i \in I, j \in J} d_{ij}$$

where d_{ij} is the distance between two objects. In our applications objects are density functions.

8.4 Suitable metrics

We now look at some metrics d_{ij} suitable for use in the above framework. We can consider in complete generality metrics defined on the space of all probability measures M defined on (Ω, B) where Ω is a polish space and B the

Borel σ-algebra; we shall let Ω be r dimensional Euclidean space. The weak or star topology is the weakest topology on M such that the map

$$F \to \int \psi \, dF$$

from M into $\mathbb{R}$ is continuous for all bounded continuous functions ψ. Losely speaking this is the smallest topology under which the above mapping is continuous. In fact we can find metrics which metrize the weak topology; the following results are taken from Huber (1981):

Lemma

For $\Omega = \mathbb{R}$, the Levy distance between two distribution functions given by

$$d_L(F,G) = \inf \{\varepsilon | \forall x \; F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon\}$$

metrizes the weak topology.

For more general Ω (remember that we require $\Omega = \mathbb{R}^r$) the somewhat complicated Prohorov metric metrises the weak topology, as does the bounded Lipschitz metric

$$d_{BL}(F,G) = \sup \left| \int \psi dF - \int \psi \, dG \right|$$

the supremum being taken over all ψ satisfying $|\psi(x) - \psi(y)| \leq d(x,y)$ where d is any distance function on Ω bounded by 1.

Of course many metrics do not metrize the weak topology, such as the Kolmogorov distance defined on the real line by

$$d_k(F,G) = \sup|F(x) - G(x)|.$$

The first two metrics have more use in a theoretical context, such metrics being the 'least discriminatory'. However we require distances which have a closed form solution, since the purpose of introducing the metrics is

to perform a clustering procedure which will enable us to reduce the computational aspect. Even the Kolmogorov metric depends on the distribution function. We shall therefore consider metrics based upon probability density functions f and g which are easier to handle. There is no loss of generality in the restriction since even in non-Normal cases we shall want to assign meaningful distributions to error terms, which in practical terms means assigning density functions.

Some discrimination measures - quantities which do not necessarily obey the triangle inequality requirement for metrics - are given by Rao (1976). Ali and Silvey (1966) give four conditions that such measures satisfy, and produce a theoretical form. Examples are, for densities f, g defined with respect to measure υ

(1) The Minkowski distance

$$\left\{\int |f(x) - g(x)|^t d\upsilon\right\}^{1/t} \qquad t \geq 1.$$

(2) Kullback-Lieber's divergence

$$J = \int (f - g) \log \frac{f}{g} \, d\upsilon.$$

This widely used measure is not a metric since it does not satisfy the triangle inequality.

(3) The Hellinger distance

$$h = \left\{\int (f^{\frac{1}{2}} - g^{\frac{1}{2}})^2 d\upsilon\right\}^{\frac{1}{2}}$$

$$= \left\{2\left(1 - \int \sqrt{fg}\, d\upsilon\right)\right\}^{\frac{1}{2}}.$$

The quantity $-\log \int \sqrt{(fg)}$ is sometimes called the Hellinger dissimilarity coefficient. This distance is a special case of Jeffrey's invariant $I_m = \int |f^{1/m} - g^{1/m}|^m \, d\upsilon$.

(4) Mahalanobis' D^2, which for two normal densities

$N_k(\underline{\mu}_i, \underline{\Sigma}_i)$ $i=1,2$ is

$$(\underline{\mu}_1 - \underline{\mu}_2)^T \left(\frac{\underline{\Sigma}_1 + \underline{\Sigma}_2}{2} \right)^{-1} (\underline{\mu}_1 - \underline{\mu}_2) \ .$$

Both (1) and (3) are true metrics as is (4) if $\underline{\Sigma}_1 = \underline{\Sigma}_2$. Although the triangle inequality is not so essential in the context of discrimination, it is important for performing a cluster analysis. We shall use the Hellinger metric; this satisfies the requirements of Ali and Silvey and has a convenient form for multivariate normal populations:

Lemma 8.1 (Matusita)

If $f_i \sim N(\underline{\mu}_i, \underline{\Sigma}_i)$ $i=1,2$ then

$$\rho \triangleq \int \sqrt{(f_1 f_2)} = \frac{|\underline{\Sigma}_1 \ \underline{\Sigma}_2|^{\frac{1}{4}}}{\left| \frac{\underline{\Sigma}_1 + \underline{\Sigma}_2}{2} \right|^{\frac{1}{2}}} \exp\left[-\tfrac{1}{4} \{ (\underline{\mu}_2 - \underline{\mu}_1)^T (\underline{\Sigma}_1 + \underline{\Sigma}_2)^{-1} (\underline{\mu}_2 - \underline{\mu}_1) \} \right] \tag{8.1}$$

Proof

From Matusita(1965)

$$\rho = \frac{|\underline{P}_1 \underline{P}_2|^{\frac{1}{4}}}{|\tfrac{1}{2}(\underline{P}_1 + \underline{P}_2)|^{\frac{1}{2}}} \exp\left[-\tfrac{1}{4} \{ \underline{\mu}_1^T \underline{P}_1 \underline{\mu}_1 + \underline{\mu}_2^T \underline{P}_2 \underline{\mu}_2 - (\underline{P}_1 \underline{\mu}_1 + \underline{P}_2 \underline{\mu}_2)^T (\underline{P}_1 + \underline{P}_2)^{-1} (\underline{P}_1 \underline{\mu}_1 + \underline{P}_2 \underline{\mu}_2) \} \right]$$

where $\underline{P}_i = \underline{\Sigma}_i^{-1}$. But

$$\underline{P}_1 \mu_1 + \underline{P}_2 \mu_2 = (\underline{P}_1 + \underline{P}_2) \underline{\mu}_1 + \underline{P}_2 (\underline{\mu}_2 - \underline{\mu}_1)$$

or $$= (\underline{P}_1 + \underline{P}_2) \underline{\mu}_2 + \underline{P}_1 (\underline{\mu}_1 - \underline{\mu}_2).$$

On substituting, the exponent becomes

$$-\tfrac{1}{4} [\mu_1^T \underline{P}_1 \mu_1 + \mu_2^T \underline{P}_2 \mu_2 - \{ (\underline{P}_1 + \underline{P}_2) \mu_1 + \underline{P}_2 (\mu_2 - \mu_1) \}^T (\underline{P}_1 + \underline{P}_2)^{-1} \ x$$

$$x \ \{ (\underline{P}_1 + \underline{P}_2) \underline{\mu}_2 + \underline{P}_1 (\underline{\mu}_1 - \underline{\mu}_2) \}]$$

which on simplifying

$$= -\tfrac{1}{4}(\underline{\mu}_2 - \underline{\mu}_1)^T \underline{P}_2(\underline{P}_1 + \underline{P}_2)^{-1}\underline{P}_1(\underline{\mu}_2 - \underline{\mu}_1)$$

$$= -\tfrac{1}{4}(\underline{\mu}_2 - \underline{\mu}_1)^T(\underline{\Sigma}_1 + \underline{\Sigma}_2)^{-1}(\underline{\mu}_2 - \underline{\mu}_1)$$

because $\underline{\Sigma}_2^{-1}(\underline{\Sigma}_1^{-1} + \underline{\Sigma}_2^{-1})^{-1}\underline{\Sigma}_1^{-1} = (\underline{\Sigma}_2 + \underline{\Sigma}_1)^{-1}$. So

$$\rho = \frac{1}{|\tfrac{1}{2}(\underline{\Sigma}_1^{-1} + \underline{\Sigma}_2^{-1})|^{\frac{1}{2}}\ |\underline{\Sigma}_1\ \underline{\Sigma}_2|^{\frac{1}{4}}} \exp[-\tfrac{1}{4}\{(\underline{\mu}_2-\underline{\mu}_1)^T(\underline{\Sigma}_1+\underline{\Sigma}_2)^{-1}(\underline{\mu}_2-\underline{\mu}_1)\}]$$

$$= \frac{|\underline{\Sigma}_1\underline{\Sigma}_2|^{\frac{1}{4}}}{|\tfrac{1}{2}(\underline{\Sigma}_1+\underline{\Sigma}_2)|^{\frac{1}{2}}} \exp\left[-\tfrac{1}{4}\{(\underline{\mu}_2 - \underline{\mu}_1)^T(\underline{\Sigma}_1 + \underline{\Sigma}_2)^{-1}(\underline{\mu}_2 - \underline{\mu}_1)\}\right]$$

which completes the proof. The last equality follows from

$$|\underline{\Sigma}_1\underline{\Sigma}_2|^{\frac{1}{2}}|\tfrac{1}{2}(\underline{\Sigma}_1^{-1} + \underline{\Sigma}_2^{-1})|^{\frac{1}{2}} = |\tfrac{1}{2}\{\underline{\Sigma}_1(\underline{\Sigma}_1^{-1} + \underline{\Sigma}_2^{-1})\underline{\Sigma}_2\}|^{\frac{1}{2}}$$

$$= |\tfrac{1}{2}(\underline{\Sigma}_2 + \underline{\Sigma}_1)|^{\frac{1}{2}}.$$

So the Hellinger distance can now be calculated using the above result in $h = \{2(1 - \rho)\}^{\frac{1}{2}}$.

The exponent is $^1/8\ D^2$ where D^2 is Mahalanobis' D^2. In the special case of equal covariance matrices $\underline{\Sigma}_1 = \underline{\Sigma}_2$, then

$$h^2 = 2(1 - \exp[-\tfrac{1}{8}\{(\underline{\mu}_2-\underline{\mu}_1)^T\underline{\Sigma}_1^{-1}(\underline{\mu}_2-\underline{\mu}_1)\}]) \qquad (8.2)$$

which is a monotone function of D^2.

Example 8.2

In the case of univariate density functions, using Lemma 8.1 with the variances σ_i for $\underline{\Sigma}_i$ gives

$$h^2 = 2\left(1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\ \exp\left[-\tfrac{1}{4}\ \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}\right]\right) \qquad (8.3)$$

thus h^2 is a function of the difference in the means of the two distributions. The exponent depends upon the

reciprocal of the sum of the variances and is premultiplied by a term depending on the ratio of these variances.

In the two special cases $\mu_1=\mu_2$, $\sigma_1=\sigma_2$ we have

$$h^2 = 2\left(1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\right) \tag{8.4}$$

and

$$h^2 = 2\left[1 - \exp\left\{-\tfrac{1}{4}\frac{(\mu_1 - \mu_2)^2}{2\sigma_1^2}\right\}\right] \tag{8.5}$$

respectively.

We have not yet specified what these measures or distances will be applied to. In many cases the natural quantity is the posterior distribution of the system vector, however we can also use the predictive distributions. Smith and Gathercole (1984) suggest the latter and are concerned that one should not combine, or cluster, distributions which are not topologically equivalent. In other words one would not wish to approximate a multimodal distribution by a unimodal one. In fact this criteria can be translated into a condition on the clustering distance as we now demonstrate in the univariate case.

Theorem 8.3

Let f_1, f_2 be any two univariate normal density functions, and let $h=h(f_1,f_2)$ be the Hellinger distance between the two distributions. Then if

$$h^2 < 2\left(1 - \frac{2e^{-\frac{1}{4}}}{(27)^{\frac{1}{4}}}\right) \tag{8.6}$$

the mixture $pf_1+(1-p)f_2$ is unimodal for all p, $0\leq p\leq 1$.

Proof

Suppose that (8.6) is satisfied, then from (8.3)

$$\sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\ \exp\left\{-\tfrac{1}{4}\,\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}\right\} > \frac{2}{(27)^{\frac{1}{4}}}\,e^{-\frac{1}{4}}\ . \qquad (8.7)$$

We claim that $\sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\ \exp\{-\frac{27}{16}\,\frac{\sigma_1^2\sigma_2^2}{(\sigma_1^2+\sigma_2^2)^2}\} \leq \frac{2\,e^{-\frac{1}{4}}}{(27)^{\frac{1}{4}}}$ (8.8)

so that from (8.7) and (8.8)

$$\exp\left\{-\tfrac{1}{4}\,\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}\right\} > \exp\left\{-\frac{27}{16}\,\frac{\sigma_1^2\sigma_2^2}{(\sigma_1^2+\sigma_2^2)^2}\right\}$$

which implies that $(\mu_1-\mu_2)^2 < \frac{27}{4}\,\frac{\sigma_1^2\sigma_2^2}{(\sigma_1^2+\sigma_2^2)}$. (8.9)

But (8.9) is a sufficient condition for the normal mixture to be unimodal (Eisenbergers result quoted by Johnson and Kotz 1970 p.89) so that the theorem is proved. It only remains to prove (8.8). Write $\sigma_1=k\sigma_2$ and consider

$$f(k) = \sqrt{\left(\frac{2k}{k^2+1}\right)}\ \exp\{-\frac{27}{16}\,\frac{k^2}{(k^2+1)^2}\} \qquad (8.10)$$

defined on $0<k<\infty$. $f(k)$ is differentiable and

$$\frac{d}{dk}\log f(k) = \frac{(1-k^2)}{2(k^2+1)}\left[\frac{1}{k} - \frac{27}{4}\,\frac{k}{(1+k^2)^2}\right] \qquad (8.11)$$

which has roots $k=1$ and $4(1+k^2)^2=27k^2$ at which points $f(k)$ takes the values $\exp(-27/64)$ and $2(27e)^{-\frac{1}{4}}$. But $f(k)\to 0$ as $k\to 0$ or $k\to\infty$ so that

$$f(k) \leq \frac{2}{(27)^{\frac{1}{4}}}\ e^{-\frac{1}{4}}$$

which establishes (8.8).

This theorem assures us that if $h^2 < 2(1 - \frac{2}{(27)^{\frac{1}{4}}}\ e^{-\frac{1}{4}})$

that is $h < 0.796$ (to three significant figures), that the mixture cannot be bimodal. A slightly stronger result can be proved if $\sigma_1 = \sigma_2$, namely

Theorem 8.4

If $\sigma_1 = \sigma_2 = \sigma$ then

$$h^2 < 2(1 - e^{-27/64}) \qquad (h<0.829) \qquad (8.12)$$

implies that the mixture is unimodal.

Proof

This follows from the proof above or directly using (8.5); if (8.12) holds then

$$\exp\left\{ - \tfrac{1}{4}\frac{(\mu_1 - \mu_2)^2}{2\sigma^2} \right\} > e^{-27/64}$$

so that

$$(\mu_1 - \mu_2)^2 < \frac{27}{8}\sigma^2 \quad \text{which is (8.9) with } \sigma_1=\sigma_2=\sigma.$$

The converse is not so relevant: if $(\mu_1-\mu_2)^2 > 8\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$ then there is a value of p for which the mixture is bimodal. One can show by arguing as above that this condition implies that

$$\sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\exp\left\{ -\tfrac{1}{4}\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2+\sigma_2^2} \right\} < \frac{1}{(2)^{\frac{1}{4}}} e^{-\frac{1}{4}}$$

in which case $h^2 > 2\{1 - 2^{-\frac{1}{4}}e^{-\frac{1}{4}}\}$

so $h > 0.830$.

8.5 Continuity Properties

Theorems 8.3 and 8.4 show that the clustering procedure applied to Gaussian mixtures with the Hellinger distance has desirable properties. We now establish certain continuity results that apply when the Kalman

updating procedure is used, but first we need to prove two subsidiary results.

Theorem 8.5

Let $\underline{X}$ be a 1 x n row vector, not identically zero and let $\underline{C}_1$, $\underline{C}_2$ be any two positive matrices. Then

$$\frac{|\underline{X}\,\underline{C}_1\,\underline{X}^T||\,\underline{X}\,\underline{C}_2\,\underline{X}^T|}{\left|\underline{X}\left(\frac{\underline{C}_1+\underline{C}_2}{2}\right)\underline{X}^T\right|} \geq \frac{|\underline{C}_1\underline{C}_2|}{\left|\frac{\underline{C}_1+\underline{C}_2}{2}\right|^2}. \tag{8.13}$$

Proof

If $\underline{R}$ is a non-singular matrix, and if we let $\underline{D}_i = \underline{R}\,\underline{C}_i\,\underline{R}^T$, $\underline{y} = \underline{X}\,\underline{R}^{-1}$ then since

$$\frac{|\underline{C}_1\underline{C}_2|}{\left|\frac{\underline{C}_1+\underline{C}_2}{2}\right|^2} = \frac{|\underline{R}\,\underline{C}_1\,\underline{R}^T||\,\underline{R}\,\underline{C}_2\,\underline{R}^T|}{\left|\underline{R}\,\left(\frac{\underline{C}_1+\underline{C}_2}{2}\right)\underline{R}^T\right|^2} = \frac{|\underline{D}_1\underline{D}_2|}{\left|\frac{\underline{D}_1+\underline{D}_2}{2}\right|^2}$$

(8.13) is equivalent to $$\frac{|\underline{y}\,\underline{D}_1\,\underline{y}^T||\,\underline{y}\,\underline{D}_2\,\underline{y}^T|}{\left|\underline{y}\left(\frac{\underline{D}_1+\underline{D}_2}{2}\right)\underline{y}^T\right|^2} \geq \frac{|\underline{D}_1\underline{D}_2|}{\left|\frac{\underline{D}_1+\underline{D}_2}{2}\right|^2}. \tag{8.14}$$

But there exists a non-singular matrix $\underline{R}$ such that $\underline{R}\,\underline{C}_1\,\underline{R}^T = \underline{I}$ and $\underline{R}\,\underline{C}_2\,\underline{R}^T = \mathrm{diag}\{\lambda_1 \ldots \lambda_n\}$ where λ_i are the eigenvalues of $\underline{C}_2$. Using this transformation in $\underline{y} = \underline{X}\,\underline{R}^{-1} = (x_1 \ldots x_n)$ say then (8.14) becomes

$$\frac{(\sum x_i^2)(\sum \lambda_i x_i^2)}{\left\{\sum \frac{(\lambda_i+1)}{2}x_i^2\right\}^2} \geq \frac{\Pi\lambda_i}{\left\{\Pi\left(\frac{\lambda_i+1}{2}\right)\right\}^2}. \tag{8.15}$$

The left-hand side in unaffected by the scaling of x_i, so that on putting $\sum x_i^2 = 1$, we require to prove

$$\left\{\frac{\sum \lambda_i x_i^2}{\sum \frac{(\lambda_i + 1)}{2} x_i^2}\right\}^2 \geq \left\{\frac{\Pi \lambda_i}{\Pi \frac{(\lambda_i + 1)}{2}}\right\}^2 . \tag{8.16}$$

Now $4z/\{(z+1)^2\}$ has a single maximum at $z = 1$, so that on any bounded interval it has a minimum on one of the boundaries. Now since $\sum x_i^2 = 1$, then $\sum \lambda_i x_i^2$ is a convex hull (on the line) and so is a bounded interval whose end points are two of the λ_i. Consequently

$$\frac{4\sum \lambda_i x_i^2}{(\sum \lambda_i x_i^2 + 1)^2} \geq \frac{4\lambda_j}{(\lambda_j + 1)^2} \quad \text{for some } \lambda_j . \tag{8.17}$$

But $4\lambda \leq (\lambda + 1)^2$

so that $$1 \geq \prod_{\substack{i=1 \\ i \neq j}}^{n} \frac{4\lambda_i}{(\lambda_i + 1)^2}$$

which combining with (8.17) gives

$$4\,\frac{\sum \lambda_i x_i^2}{(\sum \lambda_i x_i^2 + 1)^2} \geq 4\,\frac{\prod_{i=1}^{n} \lambda_i}{\prod_{i=1}^{n} (\lambda_i + 1)^2}$$

this proves (8.15) and so the theorem.

The second result that we require can be proved in a similar fashion:

Theorem 8.6

Under the conditions of Theorem 8.5, the matrix $(\underline{C}_1 + \underline{C}_2)^{-1} - \underline{X}^T\{\underline{X}(\underline{C}_1 + \underline{C}_2)\underline{X}^T\}^{-1}\underline{X}$ is positive semi-definite. (8.18)

Proof

Without loss of generality we can put $\underline{X} = \underline{X}\,\underline{R}^{-1}$, $\underline{R}$ non-singular, and invoke the simultaneous diagonalisation of quadratic forms as above, so that it is sufficient to prove

$$\begin{pmatrix} 1+\lambda_1 & & \\ & \ddots & \\ & & 1+\lambda_n \end{pmatrix}^{-1} - \underline{X}^T \left\{ \underline{X} \begin{pmatrix} 1+\lambda_1 & & \\ & \ddots & \\ & & 1+\lambda_n \end{pmatrix} \underline{X}^T \right\}^{-1} \underline{X} \geq 0 \qquad (8.19)$$

for all row vectors $\underline{X}$. But on premultiplying (8.19) by a row vector $\underline{y} = (y_1 \ldots y_n)$ and postmultiplying by $\underline{y}^T$ we have the quadratic form

$$\sum \frac{y_i^2}{1+\lambda_i} - \frac{(\sum x_i y_i)^2}{\sum x_i^2 (1+\lambda_i)}$$

which is non-negative by virtue of the Cauchy-Schwarz inequality, so that the theorem is proved. In fact the matrix of (8.18) is positive definite provided that n is larger than 1.

We are now in a position to prove certain continuity results, where the metric d refers to the Hellinger metric defined above.

Theorem 8.7

If $d(\underline{\theta}_1, \underline{\theta}_2) < \varepsilon$ then $d(\underline{X}\,\underline{\theta}_1, \underline{X}\,\underline{\theta}_2) < \varepsilon$ where $\underline{\theta}_i$ is normally distributed, $\underline{\theta}_i \sim N(\underline{m}_i, \underline{C}_i)$, and $\underline{X}$ is a fixed $1 \times n$ matrix, so $\underline{X}\underline{\theta}_i \sim N(\underline{X}\underline{m}_i, \underline{X}\,\underline{C}_i\,\underline{X}^T)$.

Proof

By definition

$$\left\{ 2 \left(1 - \frac{|\underline{C}_1 \underline{C}_2|^{\frac{1}{4}}}{|\underline{C}_1 + \underline{C}_2|^{\frac{1}{2}}} \exp\left[-\tfrac{1}{4} \left\{ (\underline{m}_2 - \underline{m}_1)^T (\underline{C}_1 + \underline{C}_2)^{-1} (\underline{m}_2 - m_1) \right\} \right] \right) \right\}^{\frac{1}{2}} < \varepsilon \qquad (8.20)$$

It follows from Theorem 8.6 that

$$\exp\left[-\tfrac{1}{4} (\underline{m}_2 - \underline{m}_1)^T \underline{X}^T \{ \underline{X} (\underline{C}_1 + \underline{C}_2) \underline{X}^T \}^{-1} \underline{X} (\underline{m}_2 - \underline{m}_1) \right]$$
$$\geq \exp\left\{ -\tfrac{1}{4} (\underline{m}_2 - \underline{m}_1)^T (\underline{C}_1 + \underline{C}_2)^{-1} (\underline{m}_2 - \underline{m}_1) \right\}$$

which combined with (8.13) gives

$$\frac{|\underline{X}\underline{C}_1\underline{X}^T|^{\frac{1}{4}}|\underline{X}\underline{C}_2\underline{X}^T|^{\frac{1}{4}}\exp\left[-\frac{1}{4}(\underline{m}_2-\underline{m}_1)^T\underline{X}^T\{\underline{X}(\underline{C}_1+\underline{C}_2)\underline{X}^T\}^{-1}(\underline{m}_2-\underline{m}_1)\right]}{\left|\underline{X}\frac{(\underline{C}_1+\underline{C}_2)}{2}\underline{X}^T\right|^{\frac{1}{4}}}$$

$$\geq \frac{|\underline{C}_1\ \underline{C}_2|^{\frac{1}{4}}}{\left|\frac{\underline{C}_1+\underline{C}_2}{2}\right|^{\frac{1}{4}}}\exp\{-\tfrac{1}{4}(\underline{m}_2-\underline{m}_1)^T(\underline{C}_1+\underline{C}_2)^{-1}(\underline{m}_2-\underline{m}_1)\}. \qquad (8.21)$$

Substituting (8.21) into (8.20) proves the theorem.

It now follows immediately that

Corollary 8.8

The mapping $N(\underline{m},\underline{C}) \to N(\underline{X}\underline{m},\underline{X}\underline{C}\underline{X}^T)$ is continuous with respect to the Hellinger metric.

If we now consider a univariate normal distribution $N(\mu,C)$ then we can show

Theorem 8.9

If V is non-negative then the mapping $N(\mu,C)\to N(\mu,C+V)$ is continuous with respect to the Hellinger metric.

Proof

If C_1, C_2 are positive then

$$\frac{(C_1+V)^{\frac{1}{4}}(C_2+V)^{\frac{1}{4}}}{(C_1+C_2+2V)^{\frac{1}{2}}} \geq \frac{(C_1C_2)^{\frac{1}{4}}}{(C_1+C_2)^{\frac{1}{2}}}. \qquad (8.22)$$

This follows from

$$(C_1+C_2)^2\{C_1C_2+V(C_1+C_2)+V^2\} \geq \{(C_1+C_2)^2+4V^2+$$
$$+4V(C_1+C_2)\}C_1C_2$$

which is true since $V(C_1+C_2)(C_1-C_2)^2+V^2(C_1-C_2)^2 > 0$.

Also $$\exp\left[-\tfrac{1}{4}\left\{\frac{(\mu_1-\mu_2)^2}{C_1+C_2+2V}\right\}\right] \geq \exp\left[-\tfrac{1}{4}\left\{\frac{(\mu_1-\mu_2)^2}{C_1+C_2}\right\}\right] \qquad (8.23)$$

Combining (8.22), (8.23) with the Hellinger distance function (8.3) for two univariate normal distribution $N(m_1,C_1),N(m_2,C_2)$ gives

$$d\{N(\mu_1,C_1+V),\ N(\mu_2,C_2+V)\} \leq d\{N(\mu_1,C_1),N(\mu_2,C_2)\}$$

which is a sufficient condition for continuity of the mapping, and the theorem is proved.

By combining the last two results, we can now prove the following important two theorems.

Theorem 8.10

The mapping $\underline{\theta}_t|y^t \to y_{t+k}|y^t$ is continuous with respect to the Hellinger metric.

Proof

From Chapter 4, if $\underline{\theta}_t \sim N(\underline{m}_t,\underline{C}_t)$ then

$$y_{t+k}|y^t \sim N(\underline{F}\,\underline{G}^k\underline{m}_t\,,\underline{F}\,\underline{G}^k\underline{C}_t(\underline{G}^T)^k\,\underline{F}^T + \overline{V})$$

where $\overline{V} = \underline{F}\,\underline{G}^{k-1}\,\underline{W}(\underline{G}^T)^{k-1}\underline{F}^T + \underline{F}^T\underline{W}\,\underline{F}^T + V$,

so that defining $\underline{X} = \underline{F}\,\underline{G}^k$, and invoking Corollary 8.8 and Theorem 8.9 proves the results since the composition of continuous functions is continuous.

This theorem shows that if $\underline{\theta}_{1t}$ and $\underline{\theta}_{2t}$ are close, then so are the predictive distributions $\underline{y}_{1t+k}|y^t$ and $\underline{y}_{2t+k}|y^t$, so that it is sufficient to see how close two system vectors $\underline{\theta}_1$ and $\underline{\theta}_2$ are. In fact with the assignments in the proof of the above theorem it follows from the proofs of Theorems 8.6 and 8.9 (using an obvious notation) that

$$d\{\ N(\underline{X}\underline{m}_1,\underline{X}\,\underline{C}_1\underline{X}^T + \overline{V})\ ,\ N(\underline{X}\,\underline{m}_2,\underline{X}\,\underline{C}_2\,\underline{X}^T+\overline{V})\ \}$$

$$\leq d\{\ N(\underline{X}\,\underline{m}_1,\underline{X}\,\underline{C}_1\,\underline{X}^T),N(\underline{X}\,\underline{m}_2\,,\,\underline{X}\,\underline{C}_2\underline{X}^T)\ \}$$

$$\leq d\{ N(\underline{m}_1, \underline{C}_1) , N(\underline{m}_2 , \underline{C}_2) \}.$$

So the predictive distributions of the observations are closer together than the system vector, that is

$$d(y_{1t+k}|y^t , y_{2t+k}|y^t) \leq d(\underline{\theta}_{1t}|y^t , \underline{\theta}_{2t}|y^t).$$

Finally we shall show that for univariate DLMs the Kalman updating procedure is continuous with respect to the Hellinger metric. In our clustering procedure, if at some time two components $N(\underline{m}_1,\underline{C}_1)$ and $N(\underline{m}_2,\underline{C}_2)$ are close then we would like posteriors corresponding to each of these components at later time stages still to be close. This means that we do not lose much by joining the two together; if this condition does not hold then the clustering procedure would be dubious. A partial answer is provided by the limit theorems of Chapter 4 which say that if the system is observable and controllable then the effects of prior knowledge decay. This means that if at time t=0 there are two priors $N(\underline{m},\underline{C}_1)$, $N(\underline{m}_2,\underline{C}_2)$ then under the same DLM $\{\underline{F}_t,\underline{G}_t,\underline{V},\underline{W}\}$ as t increases the posteriors will converge to a common limit. However in the univariate case we shall prove a more specific result:

Theorem 8.11

For a univariate DLM the Kalman updating procedure is continuous with respect to the Hellinger metric.

Proof

If at time t the system vector posterior is $N(m_i,C_i)$ then at time t+1 the posterior is $N(\overline{m}_i,\overline{C}_i)$ where

$$\overline{m}_i = gm_i + \frac{f(g^2C_i+W)(y-fgm_i)}{f^2g^2C_i + f^2W + V}$$

$$= \frac{Vm_i^*}{f^2C_i^*+V} + \frac{fyC_i^*}{f^2C_i^*+V} \tag{8.24}$$

$$\bar{C}_i = g^2C_i + W - \frac{f^2(g^2C_i+W)^2}{f^2g^2C_i+f^2W+V}$$

$$= \frac{VC_i^*}{f^2C_i+V} \tag{8.25}$$

where

$$m_i^* = gm_i \tag{8.26}$$
$$C_i^* = g^2C_i + V.$$

Without loss of generality f can be taken to be 1 (for example by putting $V=f^2V$ and $y=fy$ in (8.24), (8.25)) so that

$$\bar{m}_i = \frac{Vm_i^*}{C_i^*+V} + \frac{yC_i^*}{C_i^*+V} \tag{8.27}$$

$$\bar{C}_i = \frac{VC_i^*}{C_i^*+V}. \tag{8.28}$$

Suppose that

$$d\{ N(m_1,C_1) , N(m_2,C_2) \} < \varepsilon. \tag{8.29}$$

then from Theorem 8.10 with $F\equiv 1$

$$d\{ N(m_1^*,C_1^*) , N(m_2^*,C_2^*) \} < \varepsilon. \tag{8.30}$$

For simplicity drop the * suffix in (8.27) and (8.28). Then by the triangle inequality

$$\bar{d} \equiv d\{ N(\bar{m}_1,\bar{C}_1) , N(\bar{m}_2,\bar{C}_2) \}$$
$$\leq d\{N(\bar{m}_1,\bar{C}_1),N(\bar{m}_2,\bar{C}_1)\} + d\{N(\bar{m}_2,\bar{C}_1),N(\bar{m}_2,\bar{C}_2)\}$$

$$\leq \sqrt{2}\left[1 - \exp\{-\tfrac{1}{4}\frac{(m_1-m_2)^2V^2/(C_1+V)^2}{2VC_1/(C_1+V)}\}\right]^{\frac{1}{2}} +$$

$$+ \sqrt{2}\left[1 - \left\{2\frac{\sqrt{\left(\frac{VC_1}{C_1+V}\right)}\sqrt{\left(\frac{VC_2}{C_2+V}\right)}}{\frac{C_1}{C_1+V} + \frac{C_2}{C_2+V}}\right\}^{\frac{1}{2}}\right]^{\frac{1}{2}}. \qquad (8.31)$$

The first term is

$$\sqrt{2}\left[1 - \exp\left\{-\tfrac{1}{4}\frac{V(m_1-m_2)^2}{2C_1(C_1+V)}\right\}\right]^{\frac{1}{2}}$$

$$\leq \sqrt{2}\left[1 - \exp\left\{-\tfrac{1}{4}\frac{(m_1-m_2)^2}{2C_1}\right\}\right]^{\frac{1}{2}} \qquad (8.32)$$

because $\frac{V}{C_1+V} < 1$.

If C_1 is the maximum of C_1, C_2 then (8.32) is

$$\leq \sqrt{2}\left[1 - \exp\left\{-\tfrac{1}{4}\frac{(m_1-m_2)^2}{C_1+C_2}\right\}\right]$$

$$\leq \sqrt{2}\left[1 - \left\{\frac{2\sqrt{(C_1C_2)}}{C_1+C_2}\right\}^{\frac{1}{2}}\exp\left\{-\tfrac{1}{4}\frac{(m_1-m_2)^2}{C_1+C_2}\right\}\right]^{\frac{1}{2}}$$

$$= d\{N(m_1C_1), N(m_2,C_2)\} < \varepsilon. \qquad (8.33)$$

Now $(C_1+V)(C_2+V)(C_1+C_2)^2 \geq \{2C_1C_2 + V(C_1+C_2)\}^2 \qquad (8.34)$

because on simplifying

$$\{V(C_1+C_2) + C_1C_2\}\{(C_1+C_2)^2 - 4C_1C_2\} \geq 0.$$

(8.34) implies that $(C_1+C_2) \geq \sqrt{(C_1+V)}\sqrt{(C_2+V)}\left(\frac{C_1}{C_1+V} + \frac{C_2}{C_2+V}\right)$

giving

$$\frac{\frac{\sqrt{(C_1C_2)}}{\sqrt{\{(C_1+V)(C_2+V)\}}}}{\frac{C_1}{C_1+V} + \frac{C_2}{C_2+V}} \geq \frac{\sqrt{(C_1C_2)}}{C_1+C_2} \qquad (8.35)$$

Using (8.35) means that the second term in (8.31) is

$$\leq \sqrt{2}\left[1 - \left\{ \frac{2\sqrt{(C_1C_2)}}{C_1+C_2} \right\}^{\frac{1}{2}} \right] \leq d\{N(m_1,C_1)N(m_2,C_2)\} \leq \varepsilon \qquad (8.36)$$

from (8.29). Thus (8.31), (8.33), (8.36) imply

$$\bar{d} \leq 2\varepsilon$$

which with (8.29) gives the continuity condition, as required.

CHAPTER 9

CASE STUDIES

9.1 Introduction

We now show how some of the theory of the previous chapters could be applied in a practical context, by considering two Time Series. One is monthly U.K. chlorine production 1970 - 1982, the other is sulphuric acid production 1963 - 1982. The analyses given are not claimed to be definitive, but rather are used to illustrate by means of simple examples the way in which state-space models, Bayesian forecasting and a suitable collapsing procedure can be used with traditional techniques.

There are occasions when detailed analysis of a specific time-series is not possible, perhaps because of cost or time in a commercial setting. Consequently some kind of automatic procedure is desirable, and this is a possibility if we use the collapsing procedure of the last chapter and keep a number of models under consideration (in a similar fashion to the Class II models of Chapter 4). Traditional approaches to the problem, including many based upon Box-Jenkins models, tend to rely on a time-invariant model, fitting one model throughout the time period. But many time series do not behave in a particularly nice way, and different techniques are therefore needed.

The data are plotted in Figures 9.1 and 9.2. Certain features common to both series stand out: the extreme

FIGURE 9.1

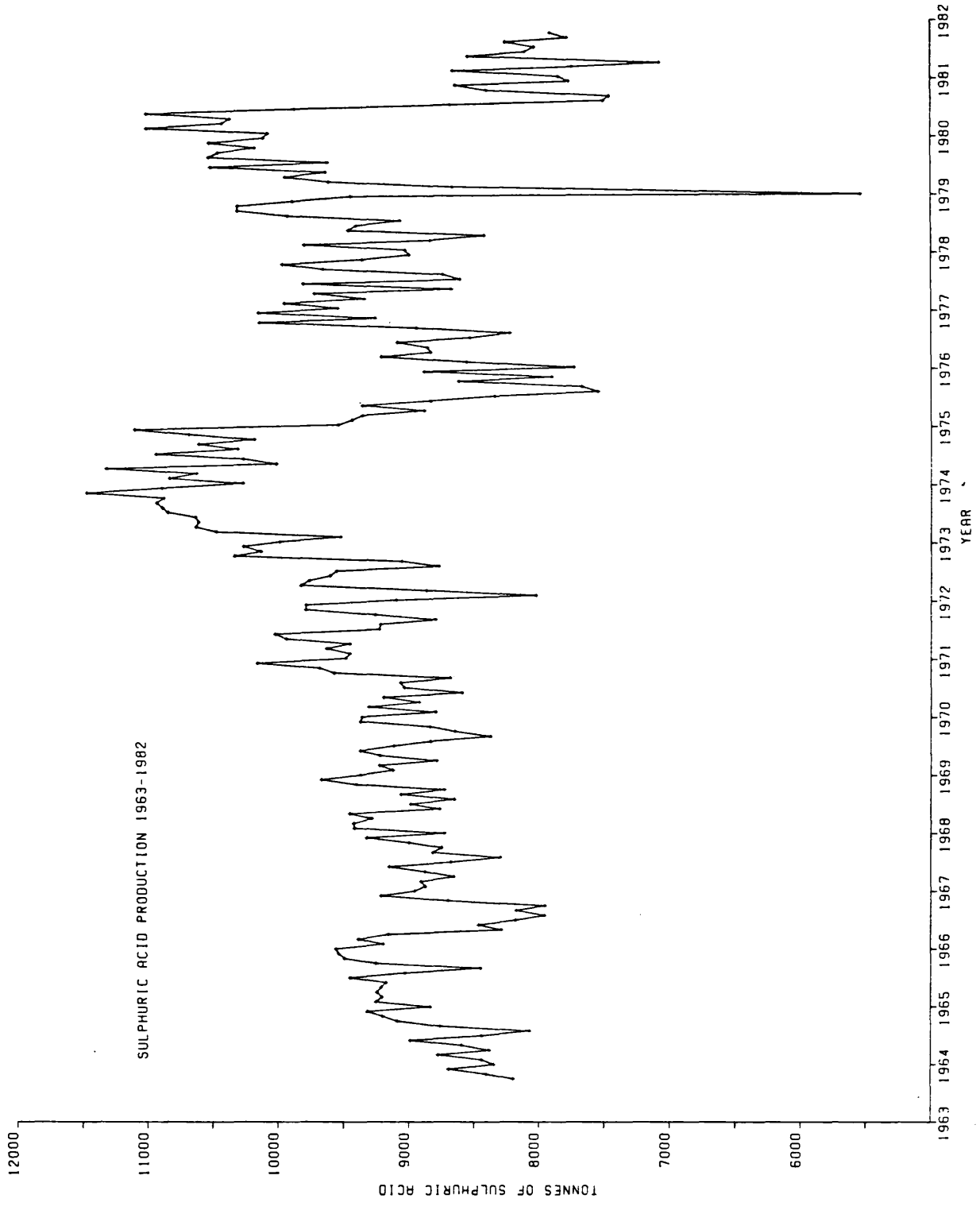
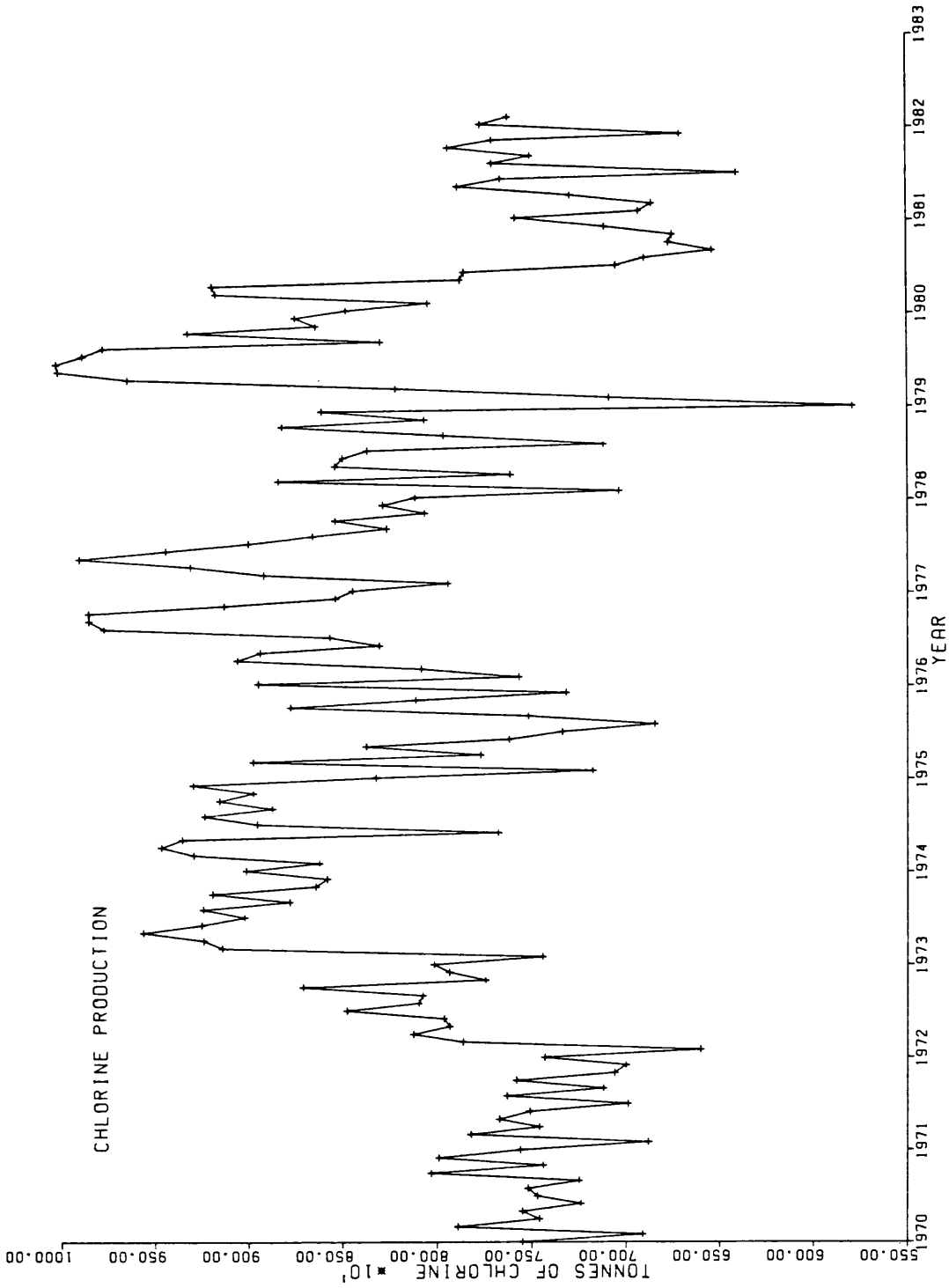

SULPHURIC ACID PRODUCTION 1963-1982
TONNES OF SULPHURIC ACID
12000
11000
10000
9000
8000
7000
6000
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
YEAR

FIGURE 9.2


CHLORINE PRODUCTION
TONNES OF CHLORINE *10'
550.00
600.00
650.00
700.00
750.00
800.00
850.00
900.00
950.00
1000.00
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
YEAR

value in January 1979 which looks like an outlier statistically, and the very different character of the data after 1974. Both of these phenomena have ready explanations: in 1979 there was a haulage strike and 1974 was the start of the oil crisis. Indeed these two series are typical of many time series which are connected in some way with the economy, characterised by a period of gradual growth up to the time of the oil crisis when suddenly all of the underlying 'steadiness' or 'stability' went. This fall was followed by a period of recovery until a second big drop around 1980 caused by the world recession.

9.2 Analysis of Sulphuric Acid Data

The sulphuric acid data has slightly more clearly defined features than the chlorine data, for example different underlying models appear more clearly. In a retrospective data analysis we might be interested in determining where a change point occurred possibly fitting different models either side of it. If we were fitting a Box-Jenkins ARIMA model then the outlier would be removed, for example by replacing with its expected value. Aternatively we might be interested in some form of trend analysis.

If on the other hand the data unfolds with time, then the 'obvious' features do not appear to be so until well after they have happened. We wish to be able to make statements about what will happen before the observations become available. In this case, as emphasised in Chapters 2 and 4 the relevant criteria are not significance tests

et al but rather how good the forecasts are.

If we only had a few observations to go on, a good starting point would be a steady model or linear growth model as described in Chapters 4, 6 and 7. This also appears to be a reasonable assumption if the correlogram of the data is examined, together with those of the differenced data - shown in Figure 9.3. These are of course based upon all of the data. The correlograms of the data transformed by taking logarithims are practically identical, so that nothing is gained by this transformation. There is perhaps a suggestion of seasonality, indicated by the somewhat larger value at lag 12, which is not surprising given the nature of the product.

In the U.K. Sulphuric acid is produced by burning sulphur rather than as a by-product of other processes as is the case in some countries, so that it is very much tied to demand. But demand for some of the products produced from sulphuric acid, such as fertilisers, can fluctuate greatly and depends upon the time of year. However the introduction of seasonal and indeed growth terms complicates an analysis designed primarily for illustration, so we shall concentrate on the simpler steady model.

The simplest steady model of Chapter 4 is

$$y_t = \theta_t + v_t \qquad (9.1)$$

$$\theta_t = \theta_{t-1} + w_t \qquad (9.2)$$

and the extended steady model introduced in Chapter 7 is

$$y_t = \theta_t + \theta_{t-1} + v_t \qquad v_t \sim N(0,V) \qquad (9.3)$$

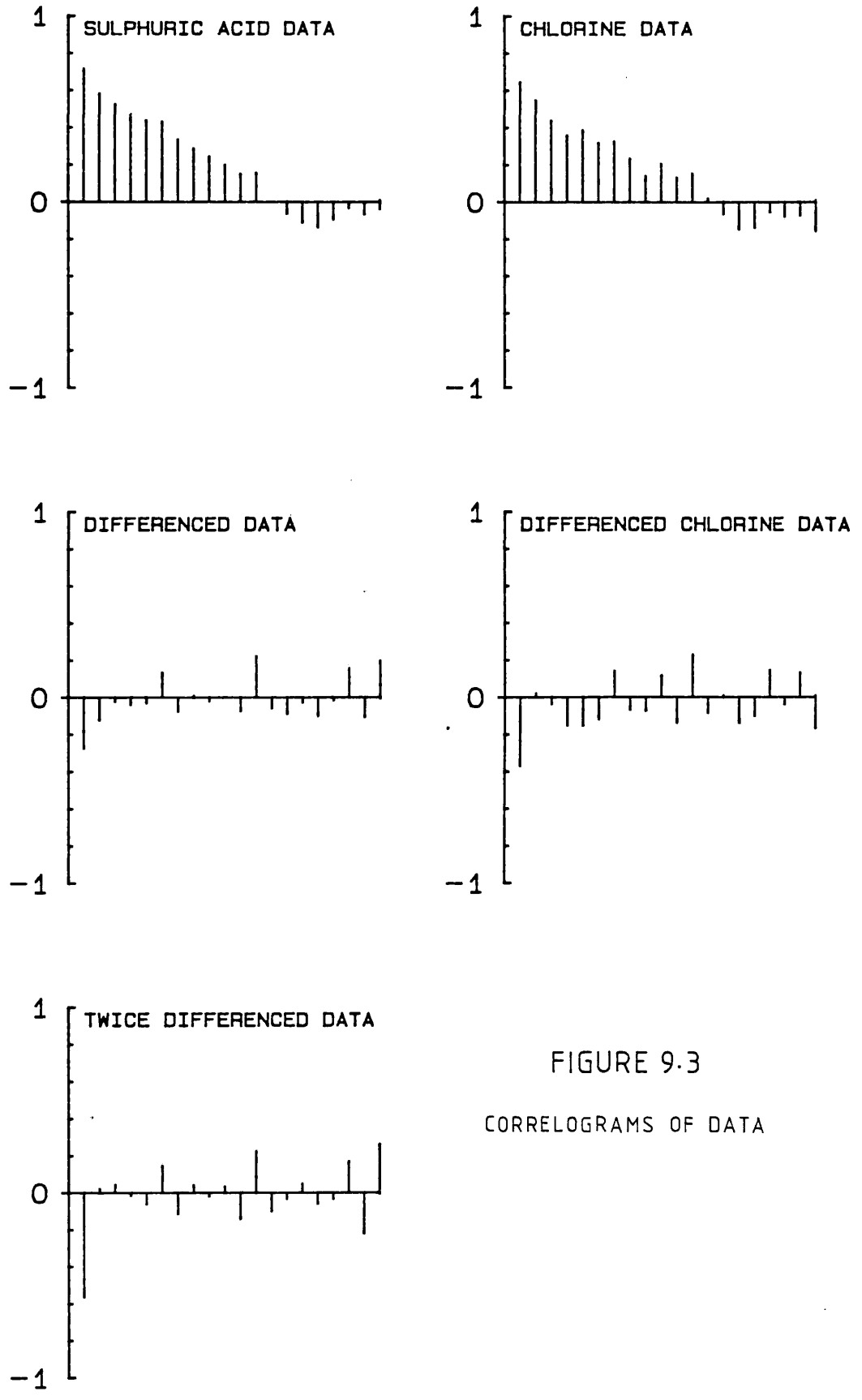


FIGURE 9.3

CORRELOGRAMS OF DATA

$$\theta_t = \theta_{t-1} + w_t \qquad w_t \sim N(0,W) \tag{9.4}$$

with the updates given in (7.79) and more generally in (7.79), (7.80) if

$$v_t \sim N(m_v, V) \tag{9.5}$$

$$w_t \sim N(m_w, W). \tag{9.6}$$

As in Chapter 7, equating the autocovariances of the differenced series $y_t - y_{t-1}$ gives

$$\gamma_0 = 2W + 2V$$

$$\gamma_1 = W - V$$

so that

$$W = (\gamma_0 + 2\gamma_1)/4 \tag{9.7}$$

$$V = (\gamma_0 - 2\gamma_1)/4. \tag{9.8}$$

An increasing level can be accounted for by putting

$$m_w = \tfrac{1}{2}(\text{mean of differenced data}). \tag{9.9}$$

Descriptive statistics for the data are shown in Table 9.1

TABLE 9.1

Sample Statistics for Sulphuric Acid Data

	Raw Data	Differenced Data
Sample size	217	216
Maximum	11439	3106
Minimum	5538	-3878
Range	5901	6984
Mean	9223.32	-1.398
Variance (unbiased)	751027.56	420251.78
Standard Deviation	866.62	648.27
Median	9200	12.5

The estimated first two autocovariances of the differenced data are

$$c_0 = 418306.17$$
$$c_1 = -113999.03$$

so that using the method of moments gives the following estimates of V and W - from (9.7), (9.8) -

$$V = 161576.07 \qquad (9.10)$$
$$W = 47577.02. \qquad (9.11)$$

The observations and one-step ahead predictors using these simple estimates of V, W are plotted in Figure 9.4. The other parameter values used were $m_V = m_W = 0$ with the prior for θ_0

$$\theta_0 \sim N(4000, 6 \times 10^6) \qquad (9.12)$$

where the large variance expresses our uncertainty of the initial mean.

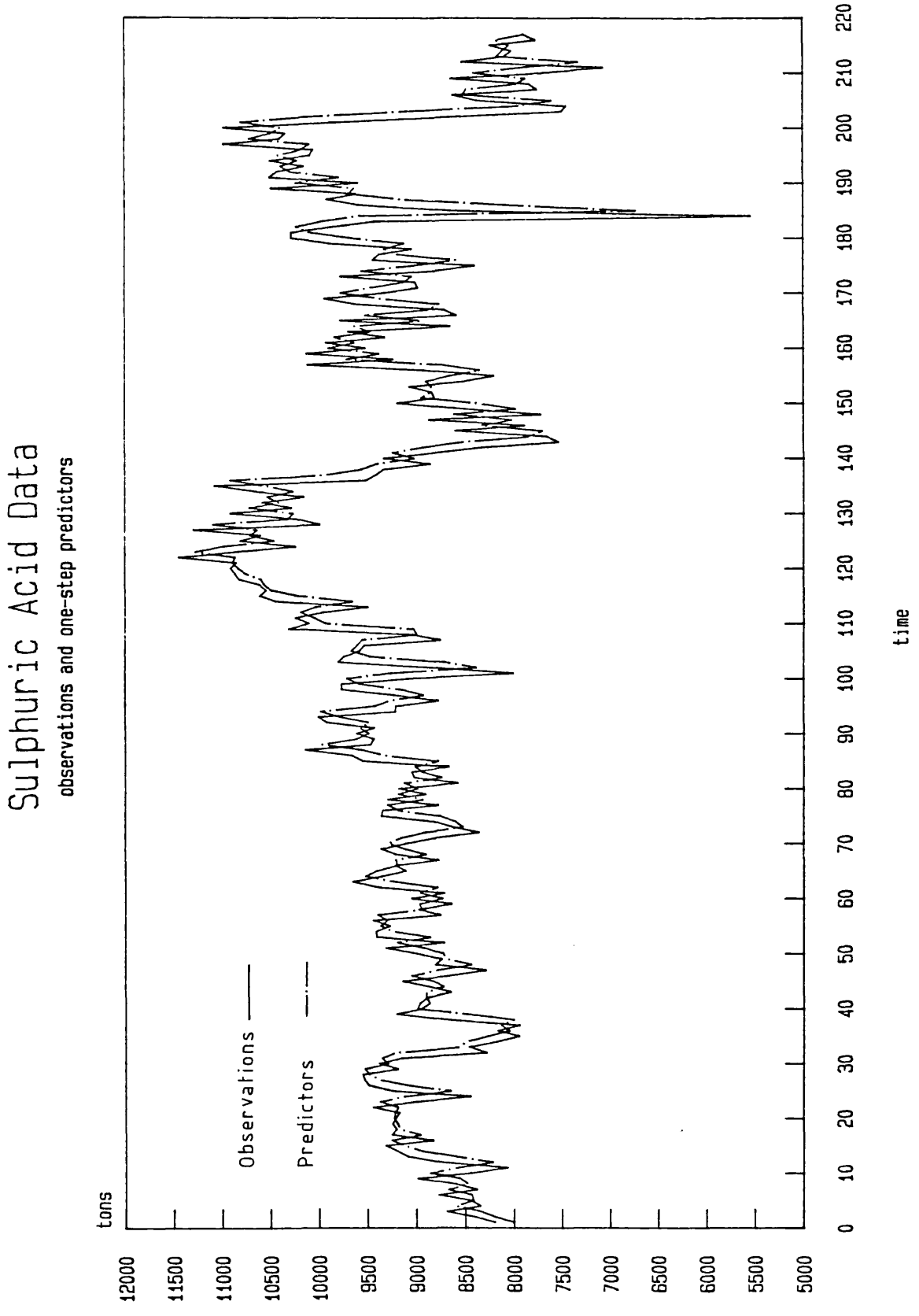
The forecasts exhibit the typical characteristics of IMA(1,1) models in that they lag the data and take some time to adjust to sudden changes in level or to outliers.

The error sum of squares for this model is 8.088×10^7.

Accounting for the slight fall in level of the differenced data by putting $m_W=(-1.398)/2 = -0.669$ gives very similar results, which is not surprising considering the small size of this term in comparison with the values of θ_t. Indeed the error sum of squares is the same to four significant figures.

For convenience we shall index the data by the integers rather than the year and month, so that t=1 corresponds to

FIGURE 9.4


Sulphuric Acid Data
observations and one-step predictors
tons
Observations
Predictors
time
12000
11500
11000
10500
10000
9500
9000
8500
8000
7500
7000
6500
6000
5500
5000
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220

October 1963, 4 to January 1964 and so up to t=217 for October 1981. Features of the data that stand out are a more or less steady evolution from t=1 to t=135, with a slight growth. A drop in level from time 136 to 143 is followed by a rising level from t=144 to 200 with a dramatic outlier at t=184. The level then drops slightly around t=201 to 202 and a steady model continues for t=203 to 217.

It is clear from (9.3) - (9.6) that a non-zero m_w term allows a deterministic rise or fall in level, which enables simple growth to be modelled rather than using the linear growth model, which involves an increase in the dimension of the system vector. With the benefit of hindsight we can perform a piecewise analysis by breaking the series up into the five parts mentioned above (apart from the outlier) and then using (9.7) to (9.9) to calculate values of V, W and m_w for each of the constituent series. The error sum of squares then reduces to 7.15×10^7, about a 10% reduction. However this piecewise approach is only really possible as a retrospective analysis, while we are interested in improved forecasting.

We now consider the period t=181 to 190 in more detail to see if it is possible to cope with the outlier in a sensible way without prior knowledge. We shall show that the use of Class II models with a collapsing procedure goes some way towards achieving this.

To begin with we use the steady model (9.3), (9.4) with parameters (9.10), (9.11) and prior (9.12). The observations and one-step ahead predictors for time t=181 are given in Table 9.2. The extreme value at time 184

drags the one-step ahead predictors down, and they remain low for the next three time-periods.

The filter behaves in such a manner because the error distributions are normal, and the normal distribution is 'outlier-resistant' - see O'Hagan (1979) - that is it tends to give outliers too much credence. A simple distribution which gives some protection against outliers is a normal mixture such as

$$v_t \sim 0.9\ N(0,V) + 0.1N(0,9V) \qquad (9.13)$$

say. So there is a small probability that the error term comes from a distribution with a much larger variance than normal, and overall the error distribution is 'heavy-tailed'. Similar distributions play an important role in the related field of robust estimation (Huber, 1981).

We now have a structure that falls into the class II models of Chapter 4. This is because using an error term (9.12) for v_t is equivalent to postulating two underlying steady models (9.3), (9.4) with in both cases W given by (9.11), but with $V_1 = V$ in (9.10) and $V_2 = 9V$. At each time-stage the transition matrix between the two models is

$$\Pi_{ij} = \begin{pmatrix} 0.9 & 0.1 \\ 0.9 & 0.1 \end{pmatrix} \qquad (9.14)$$

so that at each time stage model 1 has probability 0.9 of being in operation, independently of the past, and model 2 0.1.

On inspection it can be seen that this distribution is not really heavy-tailed enough to cope with the extreme nature of the outlier in this example, but it is one that would typically be used without such a-posteriori knowledge

of the data. This is the reason that it is used here.

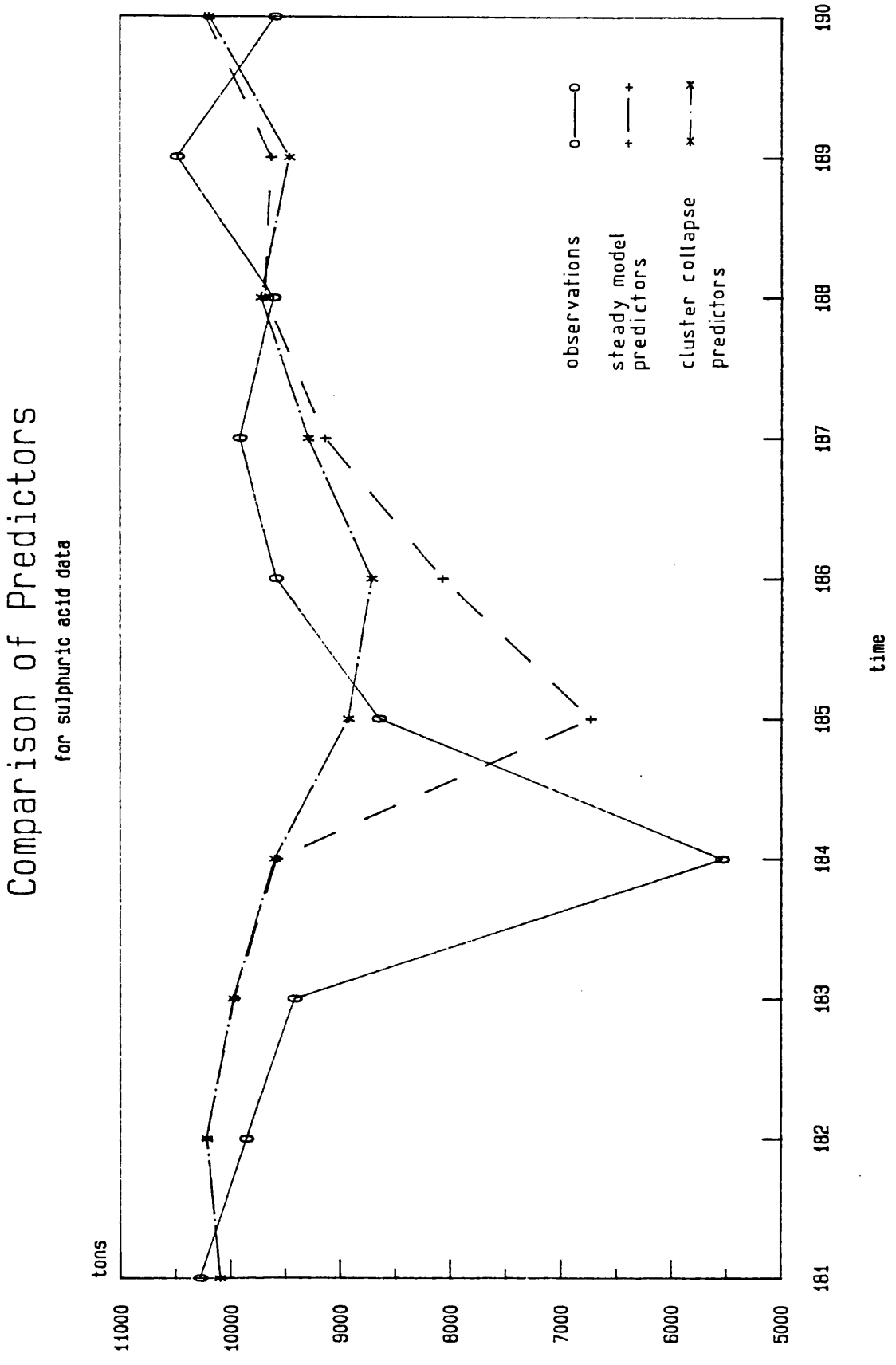
Starting from a single prior at time t=180 means that there will be 2^{10} components at time 190, and we therefore use a cluster collapsing procedure of Chapter 8 to avoid this explosion. The resulting predictors to the nearest integer are given in Table 9.2 and Figure 9.5. The predictors respond much better to the outlier than the 'ordinary' predictors, obtained using the steady model as above, in that it is largely ignored. The error sum of squares for t=181 to 190 is reduced by some 21% and the overall sum of squares is reduced by 6.4%.

TABLE 9.2

Comparison of Predictors for Time t = 181 to 190

Time	Observation	Ordinary predictor	Predictor using Cluster Collapse	Harrison -Stevens
181	10277	10096	10096	10096
182	9858	10223	10218	10218
183	9416	9966	9975	9975
184	5538	9579	9598	9599
185	8644	6736	8923	8851
186	9585	8078	8713	8695
187	9916	9138	9291	9284
188	9607	9685	9720	9717
189	10485	9630	9464	9645
190	9592	10232	10197	10197
Error sum of squares		2.4460×10^7	1.9273×10^7	1.9280×10^7

FIGURE 9.5


Comparison of Predictors
for sulphuric acid data
tons
11000
10000
9000
8000
7000
6000
5000
181
182
183
184
185
186
187
188
189
190
time
observations
steady model predictors
cluster collapse predictors

At time 180 the posterior of the system vector was taken from the the simple steady model with the parameters as above. As mentioned in the last chapter the clustering procedure used the distance between clusters defined as the greatest distance between elements in the two clusters (based on the Hellinger metric), with the nearest clusters being progressively joined..

A Harrison-Stevens Class II collapse is also included for comparison, and which gives almost identical results. This is probably because there are only two alternative models introduced at each time stage, and most of the time the posterior is dominated by one or two components. In such circumstances summing over the past histories will not be so bad, as this example illustrates. The reasons why this is not always the case were given in the last chapter and illustrated in Gathercole and Smith (1984).

We now investigate the effect of the collapsing procedure a little more closely. In §7.7 we remarked that the model (9.3), (9.4) can either be analysed as a DLM with $F = (1 \quad 1)$, $G = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$, $W = \begin{pmatrix} W_1 & 0 \\ 0 & 0 \end{pmatrix}$, or by using the modified updates given in Example 7.11. In the following we shall use the former description, so that the posterior of the system vector is effectively

$$\left(\begin{matrix} \theta_t \\ \theta_{t-1} \end{matrix} \middle| \, y^t \right).$$

At time 183, preceding the abnormal observation, the posterior is a three component mixture $\Sigma p_i N(\underline{m}_i, \underline{C}_i)$ where

the parameters are

i	p_i	$\underline{m}_i$	$\underline{C}_i$
1	.933	$\begin{pmatrix} 4789.0 \\ 4856.0 \end{pmatrix}$	$\begin{pmatrix} 43955.9 & 13212.4 \\ 13212.4 & 24240.7 \end{pmatrix}$
2	.064	$\begin{pmatrix} 4937.6 \\ 4953.2 \end{pmatrix}$	$\begin{pmatrix} 80610.5 & 36877.3 \\ 36877.3 & 39372.0 \end{pmatrix}$
3	.003	$\begin{pmatrix} 4996.0 \\ 5013.0 \end{pmatrix}$	$\begin{pmatrix} 104227.5 & 62092.0 \\ 62092.0 & 66292.4 \end{pmatrix}$

so that the most probable component dominates the mixture, this component also has the smallest covariance matrix for the system posterior, corresponding to the 'normal' size V.

The predictor at time 184 is obtained from the usual relations (as in 4.23) as

$$.933 \times 2 \times 4789.0 + .064 \times 2 \times 4937.6 + .003 \times 2 \times 4996.$$

which is 9598, the value in Table 9.2.

At time t=184, the observation is 5538. Because two models are introduced (via (9.14)) the uncollapsed posterior now has six components, with parameters

i	p_i	$\underline{m}_i$	$\underline{C}_i$
1	1×10^{-6}	$\begin{pmatrix} 3367.2 \\ 3866.4 \end{pmatrix}$	$\begin{pmatrix} 43848.9 & 13016.2 \\ 13016.2 & 23880.7 \end{pmatrix}$
2	3×10^{-6}	$\begin{pmatrix} 3234.1 \\ 3622.2 \end{pmatrix}$	$\begin{pmatrix} 46176.6 & 17286.7 \\ 17286.7 & 31715.7 \end{pmatrix}$
3	8×10^{-7}	$\begin{pmatrix} 3174.6 \\ 3513.2 \end{pmatrix}$	$\begin{pmatrix} 47098.8 & 18978.7 \\ 18978.7 & 34820.2 \end{pmatrix}$
4	0.951	$\begin{pmatrix} 4462.7 \\ 4577.3 \end{pmatrix}$	$\begin{pmatrix} 80590.3 & 36855.8 \\ 36855.8 & 39349.0 \end{pmatrix}$

5	0.047	$\begin{pmatrix} 4441.2 \\ 4554.3 \end{pmatrix}$	$\begin{pmatrix} 104288.5 & 62157.2 \\ 62157.2 & 66362.0 \end{pmatrix}$
6	0.002	$\begin{pmatrix} 4402.1 \\ 4512.5 \end{pmatrix}$	$\begin{pmatrix} 117639.0 & 76410.8 \\ 76410.8 & 81579.8 \end{pmatrix}$

The new observation means that there is an extremely small probability that the current model has a small observation error ($V_1 \equiv V$), so that the posterior probabilities of the corresponding three models are small. Note that if these posterior probabilities are not small, such as is the case when there is no alternative model, then the means of the system vector are drastically reduced from around 5000 to 3000, giving a predictor of the next observation of about 6000. This is what happends under a single steady model. With the collapsing procedure these three components are removed because of their small odds ratio, and the probability distributed amongst the remaining components.

Using the Hellinger distance function gives the distance between the three models as

	4	5	6
4	0	0.1896	0.2841
5		0	0.1057
6			0

It follows that components 4 and 6 are the furthest apart, and 5 and 6 the closest. Components 5 and 6 are closer together than 4 and 5 because although the means of the latter two are more similar, the covariance matrices are not.

The cluster cut-off distance was set at 0.2 in the above analysis, mainly for pragmatic reasons: it

reduced the mumber of models reasonably, and also in much less than the upper bounds discussed in the last chapter. As we have already mentioned, the choice of this figure needs further investigation. With this assumption components 5 and 6 are clustered together, but 4 is not further added because using furthest neighbour distances, the distance of 4 from {5,6} is 0.2841. This results in the collapsed posterior

$$0.951\ N\left[\begin{pmatrix}4462.7\\4577.3\end{pmatrix},\begin{pmatrix}80590.3 & 36855.8\\ & 39349.0\end{pmatrix}\right]$$

$$+\ 0.049\ N\left[\begin{pmatrix}4439.3\\4552.3\end{pmatrix},\begin{pmatrix}104992.11 & 62908.4\\62908.4 & 67164.0\end{pmatrix}\right]$$

The second component comes from joining 5 and 6 where the 5th component has the predominant effect in the mixture because the cluster centre is formed by weighting the points according to their probabilities.

Table 9.3 shows how the number of models involved at each stage alters

TABLE 9.3 : Summary of posterior component evolution

Time	Number of models after collapse	probabilities				Most probable mean & variance of θ	
181	2	.098	.052			5111.6	43838
182	3	.943	.054	.003		4981.9	43985
183	3	.933	.064	.003		4789.0	43956
184	2	.951	.049			4462.7	80590
185	2	.940	.060			4351.7	46223
186	4	.826	.073	.094	.006	4665.0	44043
187	4	.887	.042	.066	.005	4873.3	43933
188	4	.892	.057	.048	.003	4823.8	43859
189	4	.848	.056	.092	.005	5117.9	43854
190	4	.806	.121	.065	.007	4892.0	43856

The table also illustrates the effect of the collapsing procedure: nearly all the time the most probable component is the result of assuming that $V_1 = V$, giving a posterior variance for θ_t of the order of 4×10^5. However when the extremely low observation appears, these components are filtered out by the collapsing procedure, leaving posterior variances of the order of 8×10^5 or more, that is the term $V_2 = 9V$ comes into play. As the next observation is close to the predicted means, the 'normal' system variance again becomes the dominant term and stays that way. Thereafter the number of collapsed models remains at 4, a number greater than 1 partly because a better model would be more complex than a steady model.

9.3 Measures of Accuracy

The Kalman filter produces a predictive distribution, so that as we have mentioned before, using a summary statistic - such as the mean - to represent this distribution involves some loss of accuracy. However when point forecasts are required, some measure of their accuracy is required. Much of the discussion in Chapter 4 on loss functions can therefore be applied, since loss functions are appropriate measures of accuracy. More specifically, if we have a single forecast f of a point p, both possibly multidimensional, then a loss function $L(f,p)$ which maps onto the reals is a suitable measure. Typically we look for a function of the form $L(f - p)$, with $L(x) \geq 0$, and $L = 0$ because $f = p$ is a 'perfect' forecast.

In a practical situation it is unlikely that L will be a symmetric function because there are many reasons why

underestimating or overestimating will produce different consequences. It is likely that only in the abstract context of estimation will we be unconcerned as to the direction of the error. Nevertheless, in making comparisons between different methods it is more helpful to use symmetric loss function rather than asymmetric ones.

The classical measure of error is the sqared loss function

$$L(y, \hat{y}) = (y - \hat{y})^2$$

which has a number of drawbacks, some of which are touched upon in Harrison and Stevens (1976). Another choice when we have normal distributions is to use the loss function 'conjugate' to the normal density function (Lindley (1976)),

$$L(y,\hat{y}) = h[\,1 - \exp\{-\tfrac{1}{2k}(y - \hat{y})^2\}\,] \quad . \qquad (9.15)$$

Without loss of generality we take $h \equiv 1$, so that

$$L(y,\hat{y}) = 1 - \exp\{-\tfrac{1}{2k}(y - \hat{y})^2\}.$$

We shall only consider the one-step ahead predictor, but this is not the only possibility; for example we might be interested in forecasts at different lead times, such as a year ahead (lag 12).

Another possible measure of accuracy given in Harrison-Steven (1976) is to consider cumulative forecasts, for six months say. In effect this looks at $L(\sum y_t, \sum \hat{y}_t)$ rather than $\sum L(y_t,\hat{y}_t)$, which in general will be different.

9.4 Chlorine Data Analysis

Some results of using different forecasting techniques to-gether with measures of performance introduced in the previous section are presented in Table 9.4. The figures have been scaled across the rows (which is equivalent to altering the constant premultiplier h of the loss function (9.15)) to give a loss of 100 for Method 1; this enables the different procedures to be compared more easily. The methods used were as follows:

First, in Method 1, the classical EWMA approach was used, fitting an IMA(1,1) model with moving average parameter β estimated through the iterative likelihood equation (2.24), giving the estimate $\beta = -0.5253..$.

Method 2 used the steady DLM (9.3) - (9.6) with the free parameters estimated from the method of moments via (9.7) - (9.9) giving

$$W = 36392 \qquad (9.16)$$

$$V = 242822. \qquad (9.17)$$

Not surprisingly the mean square error is larger than for the first method - indeed Example 2.4 demonstrated that the maximum likelihood equation is equivalent to minimising this quantity. However for some of the conjugacy parameters k this Kalman filter gives slightly better results. The conjugate loss function behaves differently to the mean square error criterion, as this and the following examples illustrate.

Methods 1 and 2 are both retrospective - they presuppose knowledge of the whole (or at least part) of the Time Series. In practice we would be uncertain as to the

TABLE 9.4

Comparison of methods and loss functions

Parm. k	Method 1	1	2	3	4	5	6
		← scaled →					
1	144.76	100	100.17	100.02	99.48	100.17	100.15
10	144.13	100	100.59	100.01	99.91	100.35	100.12
500	141.76	100	98.69	99.16	98.65	99.07	98.66
1000	139.55	100	98.54	98.99	98.72	98.91	98.75
5000	129.03	100	98.30	98.51	98.85	98.46	98.79
10^4	122.03	100	98.38	98.53	98.94	98.50	98.93
5×10^6	6.12	100	100.47	100.74	99.89	100.81	99.56
Squared loss	6.665 $\times10^7$	100	100.60	100.89	99.73	100.95	99.37

Method 1: Classical EWMA from IMA (1,1) model.

2: Augmented steady DLM.

3: Class I procedure with a grid of values for V, 1 for W.

4: Class II procedure with collapsing $V \sim 0.9N(0,V) + 0.1N(0,9V)$.

5: Class II procedure with grid of values of W and V.

6: Class II with collapsing, V and W as in 5.

The parameter k refers to the loss function (9.15).

values of V and W for at least the first few time points. As a first step towards modelling this, Method 3 assumes W known with value (9.16), but V unknown and described by an equally spaced grid of values as follows:

$$V_1 = 2 \times 10^5, \quad V_5 = V \text{ from (9,17)}, \quad V_9 = 2.85645 \times 10^5$$
$$\text{and } V_i \text{ such that } V_1, V_2 \ldots V_9 \text{ equally spaced} \tag{9.18}$$

with the prior probabilities

$$p_i = i/25 \qquad i \leq 5$$
$$p_i = p_{10-i} \qquad 6 \leq i \leq 9 \tag{9.19}$$

where $p_i = \text{Prob}\{V=V_i\}$.

V_5 is therefore the most probable component initially.

The Harrison-Stevens Class I model was then used with these assumptions, as described in Chapter 4. As can be seen from the table, overall this gives a similar perfomance to Method 2, with slightly lower loss for k=1 and 10, slightly higher otherwise.

In percentage terms the differences are small, indeed for all the various methods the variations across the table are small. In part this is because the data would be better described by a more sophisticated model - the steady model is not wholly appropriate. However even the small differences show that it is possible to use more realistic models than those which presuppose the data and do as well, or even slighlty better than a retrospective analysis.

Method 4 was the jump method used in the analysis of the sulphuric acid data. That is with W as in Method 2, given by (9.17), we allow the error term v_t to take the

value V from (9.16) with probability 0.9, and 9V with probability 0.1, or

$$v_t \sim 0.9\ N(0\ ,\ V) + 0.1\ N(0\ ,\ 9V)$$

which is a heavy-tailed distribution robust against outliers. The collapsing procedure of Chapter 8 was used with the parameters set as in §9.3.

This method gave better results than the classical EWMA for all the conjugate paramters, and also for squared loss. Again overall the differences are small because although this error distribution gives much improved estimates around the outlier, which considerably reduces the squared error locally, the introduction of two models at each time stage gives slighlty poorer estimates if there are no outliers or changes in level. Comparing with §9.3 it can be seen that the improvement that this method gives is much less marked for the chlorine data than for the sulphuric acid data.

Methods 5 and 6 extend Method 3 by putting a grid of values on W as well as V, with 9 different values for W chosen in an analagous way to that given above for V in (9.18), (9.19) with in this case $W_5 = W$ from (9.17). Method 5 used the Class I procedure, so that one of the 81 models is assumed to hold over the entire time period, whereas Method 6 used a Class II model with the collapsing procedure of Chapter 8. This is equivalent to assuming a normal error distribution of 9 components for v_t and similarly for w_t which holds at each time stage.

In terms of mean square error, and for nearly all the parameters of the loss function, Method 6 which uses

the collapsing procedure gave slightly better results than Method 5. In fact in terms of mean square error this approach gave the best improvement over the maximum likelihood approach, which underlines the point made earlier that it is possible to use models (DLMs with a collapsing procedure) that do not require the data to be known in advance and obtain good results.

The above comparisons also illustrate that using different criteria when assessing forecasting performance can favour different models according to the criterion chosen. In the results given, it can be seen that the bounded loss function behaves differently to the usual squared loss function.

9.5 Further Analysis

For the reasons stated it is difficult to draw strong conclusions about the relative merits of the different methods from analysing the chlorine data. It is conjectured that when parameter values in DLMs are unknown it is better to use a Class II collapsing procedure rather than a Class I procedure (unless it is required to identify a particular parameter value). This needs to be substantiated by further research, and the question of how many grid-points to chose and over what range of values also need investigation. Certainly the sulphuric acid data shows the advantage of using outlier resistant error distributions when there is the possibility of outliers or changes in level.

More sophisticated analyses of either Time Series would

involve more complicated models, such as those involving growth terms and possibly seasonal factors. Also the two series are not independent, since at the very least they are related via the underlying economic climate, and it would be possible to perform a bivariate analysis, so that the observation vector in the appropriate DLM has dimension 2. A preliminary investigation of the possibility of using a bivariate analysis is to assume one series known and then use the information fron this series to interact with the other (asumed unkown) to improve the forecasts. Initial attempts at this on the chlorine data, using the interactions described in Chapter 7 on the steady model, suggest that it is possible to reduce the mean square error by over 15%. Of course a true bivariate forecasting model dose not assume one series known.

CHAPTER 10

SUMMARY

This thesis examines the use of state-space models in Time Series analysis, particularly those of the form

$$\underline{y}_t = \underline{F}\,\underline{\theta}_t + \underline{v}_t$$

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

which relate the observations $\underline{y}_t$ to a state vector $\underline{\theta}_t$; these are termed Dynamic Linear Models (DLMs) by Harrison-Stevens (1976). They coined the phrase 'Bayesian forecasting', for these models are naturally amenable to Bayesian inference. Indeed, analysis by the Kalman filter can be viewed as successively updating the posterior for $\underline{\theta}_t$ by incorporating new information as it arrives. Forecasts or predictors are obtained from the appropriate predictive distribution function by using a suitable loss function. The background is summarised in Chapters 2 to 4, which includes certain concepts from control theory that we use.

The usual assumption in the above DLM is that $\underline{v}_t$, $\underline{w}_t$ are independently normally distributed, $v_t \sim N(0,\underline{V})$, $\underline{w}_t \sim N(0,\underline{W})$ say. Chapter 5 considers the possibility of relaxing the assumptions by allowing distributions other than Gaussian ones. If the additive description is retained then handling other distributions soon becomes intractable because of the convolutions involved. This is true even for the simplest steady model

$$y_t = \theta_t + v_t$$

$$\theta_t = \theta_{t-1} + w_t$$

unless the error terms have a stable law, so an alternative description is needed if we are to be able to use non-normal distributions.

Using the normal steady model as a starting point, Smith (1979) proposes a class of steady evolutions. This specifies the conditional distribution $f(y_t|\theta_t)$, and the evolution of $\theta_t|y^{t-1}$ to $\theta_t|y^t$ by $p(\theta_t|y^t) \propto p(\theta_t|y^{t-1})^{k_t}$ for some constant k_t. The implications for the predictors of such a system are discussed in Chapter 5; $f(y_t|\theta_t)$ is taken to be a member of the exponential family with $p(\theta_t|y^t)$ the appropriate conjugate. Theoretical results concerning the predictive distributions of the observations and the system vector are derived, and conditions under which the evolution corresponds to an invariant transition density proved.

With this scenario the nature of the forecasts depends very much upon both the model chosen and the loss function used. The properties of the normal model are constant forecasts at all lead times, with the uncertainty associated with the forecasts increasing with the lead time. Examples are given which show that it is possible to construct non-normal models that under symmetric loss functions have one, both or neither of these properties.

Chapter 6 explores the relationships between DLMs (with normal errors) analysed by the Kalman filter, and ARIMA models; comprehensive equivalence theorems are proved, where equivalence is taken to mean that the forecasts of the two descriptions are identical for all lead times. Firstly, DLMs are shown to be predictor equivalent to their

observable subsystems, so that we only need to consider observable DLMs. If a DLM $\{\underline{F}, \underline{G}\}$ is observable then $\underline{G}$ is non-derogatory, that is $\underline{G}$ is similar to a companion matrix and has minimal polynomial identical to its characteristic polynomial. If $\underline{G}$ is of order n, then this is $\lambda^n + \sum_{i=1}^{n} \phi_i \lambda^{n-i}$ say. Let s be the largest integer such that ϕ_s is non-zero, and define r the 'system shift' to be the maximum of 0 and n-s-1. Two cases arise, either $\underline{G}$ is non-singular, so that s=n and $\underline{G}$ has rank n, r=0, or $\underline{G}$ is singular with rank n-1 and system shift $r = n-s-1 \geq 0$.

A DLM with non-derogatory matrix $\underline{G}$ which is in the steady state, so that the Kalman gain vector $\underline{A}_t$ has converged to $\underline{A}$, is shown to be predictor equivalent to the ARMA model

$$y_t + \phi_1 y_{t-1} + \ldots + \phi_s y_{t-s} = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_{r+s} \varepsilon_{t-r-s}$$

where the ϕ_i's are obtained from the characteristic polynomial and the β_j's depend upon the ϕ_i's and $\underline{A}$ through $\underline{F}\,\underline{G}^j\,\underline{A}$. In fact this result is independent of assumptions on the ϕ_i's and so includes non-stationary models as well, in particular the ARIMA models. Once the autoregressive parameters are obtained, then it is possible to derive the β_i's by considering the covariance properties of the DLM, which gives the same values provided that certain mild restrictions hold.

The stability conditions for the DLM are shown to coincide with the invertibility conditions for the equivalent ARIMA process; these will be satisfied if the DLM is observable and controllable. The equivalence theorems are

extended to incorporate time-varying DLMs in which $\underline{F}$, $\underline{G}$ and $\underline{A}_t$ can vary in time - in this instance the parameters ϕ_i and β_j are also time-dependent.

An examination of the structural properties of DLMs is started in Chapter 6 and extended in Chapter 7. In the case of zero system shift, r=0, $\underline{G}$ can be singular or non-singular. The latter can place severe restrictions on the parameters $\beta_1 \ldots \beta_s$. This surprising result says that a singular matrix $\underline{G}$ is preferable for zero system shift, where the order of the system vector is one larger than it need be, namely n+1 rather than n. For example with the steady model, an unrestricted system vector is of order 2 rather than 1, as pointed out by Godolphin and Stone (1980).

For an observable univariate DLM with constant matrices $\underline{F}$ and $\underline{G}$, when the error variances V and $\underline{W}$ are allowed to vary, an image is traced out in the invertibility space of the equivalent ARMA model. V must be positive and $\underline{W}$ positive semi-definite, and if this is the domain of the mapping then, provided the system is observable and has large enough order (n = r + s + 1), it is possible to cover the entire invertibility region. This is proved in Chapter 7. The invertibility region can be pictured in the parameter space of $(\beta_1 \ldots \beta_{r+s})$ or equivalently in the correlation space $(\rho_1 \ldots \rho_{r+s})$, the autocorrelations of $y_t + \phi_1 y_{t-1} \ldots + \phi_s y_{t-s}$.

In a practical context only the diagonal terms in $\underline{W}$ can be realistically updated, unless an estimation procedure is applied. With $\underline{W}$ restricted to a diagonal matrix (or equivalently to $\underline{B}\, W_d\, \underline{B}^T$, with $\underline{W}_d$ diagonal and $\underline{B}$ fixed), the

image in the autocorrelation space is a convex hull of its vertices. Consequently unless the invertible region of the ARIMA model is a convex polytope, it will not be possible to cover the full region with such practically attractive DLMs.

These results are illustrated by considering ARIMA (p,d,q,) models with $p+d=1$ or 2, $q \leq p+d$. For first order models, by choosing a system matrix of order 2 it is possible to cover the full invertibility region with a DLM having a diagonal covariance matrix. However the appropriate region for second-order models, such as ARMA (2,2), cannot be so covered, since the region is not a convex hull of a finite set of points. Either an infinitely large DLM has to be used, or a DLM with a non-diagonal matrix $\underline{W}$ whose elements can vary. However, it is possible to cover a larger part of the region than that achieved by models advocated by Harrison-Stevens, which suffer from the defect of too small a dimension. A summary is provided which gives second-order ARIMA models, the equivalent Harrison-Stevens DLMs, unrestricted equivalent DLMs and suitable practical (possibly restricted) DLMs.

An alternative way of writing an unrestricted DLM equivalent to a steady model is

$$y_t = \theta_t + \theta_{t-1} + v_t$$

$$\theta_t = \theta_{t-1} + w_t$$

which has the attraction of a single system parameter, rather than a vector of order 2 when it is expressed in standard DLM form. The so-called 'Markov-Polynomial' models can be also be extended and written in this 'augmented' form.

The analysis of these models is discussed in Chapter 7; it is shown that it is possible to derive a slightly different form of the Kalman filter for such 'augmented' models.

Chapter 8 focusses attention on the Multiprocess models of Harrison-Stevens, looking at alternative ways of collapsing the components of the normal mixture of the posterior at each time stage. A class of clustering collapsing procedures is developed, where the distance between normal components is used together with a clustering procedure to form clusters of the components. The cluster centres can then be used to summarise the information in the cluster, effecting a reduction in the number of components.

The Hellinger metric is chosen as a suitable metric and certain desirable properties follow, for example if the distance between two unimodal densities is below a certain threshold then the mixture is unimodal. This means that if we use say agglomerative clustering and stop forming clusters before this threshold is reached, then we would not approximate a bimodal distribution by a unimodal one, which is important not least from a decision theoretic standpoint.

Certain continuity properties are then proved with respect to the Kalman filter, which show that the clustering is sensible. That is, if two components of the system posterior are close then so are the corresponding components of the predictive distribution, and successive posteriors remain close under the influence of the filter. This last point, which says that little is lost by combining components that are close is only proved for a univariate DLM with univariate system matrix. The effect of the clustering procedure chosen, and the choice of 'cut-off' distance if

using agglomerative clustering needs further research.

Chapter 9 applies some of the preceding theory, particularly of Chapters 7 and 8, to two real Time Series, namely monthly sulphuric acid production in the U.K. and monthly chlorine production from 1963-1982. These series exhibit characteristics typical of much economic-related data of the past two decades, such as a period of steady growth followed by a dramatic fall caused by the oil crisis of the mid-seventies, with the ensuing recovery punctuated by a number of depressions. In these circumstances a single time-invariant model is unlikely to produce optimal forecasts, but there are situations where forecasts need to be made automatically, without the direct intervention of an experienced statistician.

It is possible to use a semi-automatic procedure, by using the Multiprocess procedure of Harrison-Stevens, together with a suitable collapsing procedure. Ideally one would have different types of DLMs with the Multiprocess models, however for illustrative purposes the augmented steady model is used, with error variances that can vary. Using the collapsing procedure introduced in Chapter 8, it is possible to improve on the performance of the traditional EWMA of the IMA (1,1) model. The performance is markedly improved in the region of an 'outlier', to which the EWMA is very sensitive, enabling outliers to be automatically treated statistically. It is also shown that the use of different measures of performance - other than the usual squared-error criterion - can favour different methods.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I.A. (1965). Handbook of Mathematical Functions. New York: Dover.

AKAIKE, H. (1974). Markovian representation of stochastic processes, and its application to the analysis of autoregressive-moving average processes. Ann. Inst. Statist. Math., 26, 363-383.

ALI, S.M. and SILVEY, S.D. (1966). A general class of coefficients of divergence of one distribution from another. J.R. Statist. Soc. B, 28, 131-142.

BARNETT, S. (1975). Introduction to Mathematical Control Theory. Oxford: University Press.

BATHER, J.A. (1965). Invariant conditional distributions for Bayesian inference. Ann. Math. Statist., B, 36, 829-846.

BERNARDO, J.M. (1979). Reference posterior distributions for Bayesian inference. J.R. Statist. Soc. B, 41, 113-147.

BIRKHOFF, G. and MACLANE, S. (1965). A Survey of Modern Algebra. New York: MacMillan.

BOX, G.E.P. and JENKINS, G.M. (1970). Time Series Analysis, Forecasting and Control. (2nd edition: 1976) San Francisco: Holden Day.

BOX, G.E.P. and TIAO, G.C. (1973). Bayesian Inference in Statistics. Reading, Mass: Addison-Wesley.

BROWN, R.G. (1959). Statistical Forecasting for Inventroy Control. New York: McGraw-Hill.

CHATFIELD, C. (1977). Some recent developments in time series analysis. J.R. Statist. Soc. A, 140, 492-510.

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.

COX, D.R. (1981). Statistical analysis of time series: some recent developments. Scand. J. Statist., 8, 93-115.

COX, D.R. and HINKLEY, D.V. (1974). Theoretical Statistics. London: Chapman Hall.

DE GROOT, M.H. (1970). Optimal Statistical Decisions. New York: McGraw-Hill.

FELLER, W. (1970). Introduction to Probability Theory. New York: Wiley.

FERGUSON, T.S. (1967). Mathematical Statistics: a Decision Theoretic Approach. New York: Academic Press.

GATHERCOLE, D. and SMITH J.Q. (1984). A dynamic forecasting model for a general class of discontinuous time series. Time Series Analysis: Theory and Practice 5, (O.D. Anderson. ed). North-Holland.

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. (1976). 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. 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.

HAGGAN, V. and OZAKI, T. (1981). Modelling non-linear random vibrations using an amplitude dependent auto-regressive time series model. Biometrika, 68, 189-196.

HANNAN, E.J. (1960). Time Series Analysis. London: Chapman-Hall.

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.

HUBER, P.J. (1981). Robust Statistics. New York: Wiley.

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.

JEFFREYS, H. (1961). Theory of Probability. Oxford: University Press.

JENKINS, G.M. (1979). Modelling and Forecasting Time Series. Lancaster: G.J.P.

JOHNSON, N.L. and KOTZ, S. (1970). Continuous Univariate Distributions - 1. Boston: Houghton Mifflin.

JURY, E.I. (1964). Theory and Application of the z-Transform Method. New York: Wiley.

KALMAN, R.E. (1963a). New methods in Weiner filtering theory. In Proc. 1st Symp. on Eng. Applications of Random Function Theory and Probability Theory. (J.L. Bogdanoff and F. Kozin, eds). New York: Wiley.

KALMAN, R.E. (1963b). Mathematical description of linear dynamical systems. J. SIAM Control A, 1, 152-192.

KENDALL, M. (1976). Time Series. 2nd edition. London: Griffin.

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. New York: Holt, Rinehart and Winston.

KWAKERNAAK, H. and SIVAN, R. (1972). Linear Optimal Control Systems. New York: Wiley.

LAWRENCE, A.J. and LEWIS, P.A.W. (1980). The exponential autoregressive-moving average EARMA(p,q) process. J.R. Statist. Soc. B, 42, 150-161.

LINDORFF, D.P. (1965). Theory of Sampled-Data Control Systems. New York: Wiley.

LINDLEY, D.V. (1971). Bayesian Statistics, a Review. SIAM, Philadelphia.

LINDLEY, D.V. (1976). A class of utility functions. Ann. Statist., 4, 1-10.

MANSKI, C.F. (1981). Learning and decision making when subjective probabilities have subjective domains. Ann. Statist., 9, 59-65.

MATUSITA, S. (1965). A distance and related statistics in multivariate analysis. In Proc. 1st Int-Symp. on Multivariate Analysis, Dayton Ohio, 187-200.

MCKENZIE, E. (1976). A comparison of standard forecasting systems with the Box-Jenkins approach. The Statistician, 23, 107-116.

MUTH, J.R. (1960). Optimal properties of exponentially weighted forecasts. J. Amer. Statist. Ass., 55, 299-306.

O'HAGAN, A. (1979). On outlier rejection phenomena in Bayes inference. J.R. Statist. Soc. B, 41, 358-367.

PRIESTLEY, M.B. (1980). State dependent models; a general approach to non-linear time series analysis. J. Time Series Analysis 1, 47-71.

RAIFFA, H. and SCHLAIFFER, R. (1961). Applied Statistical Decision Theory. Boston: Harvard Business School.

RAO C.R. (1976). Cluster analysis applied to a study of race mixture in human populations. Classification and Clustering, (J. Van Ryzin, ed). New York: Academic Press.

SORENSON, H.W. and ALSPACH, D.L. (1971). Recursive Bayesian estimation using Gaussian sums. Automatica, 7, 465-479.

SMITH, J.Q. (1978). Problems in Bayesian statistics relating to discontinuous phenomena, catastrophe theory and forecasting. Ph.D. Thesis, University of Warwick.

SMITH, J.Q. (1979). A generalisation of the steady forecasting model. J.R. Statist. Soc. B, 41, 375-387.

SMITH, J.Q. (1980). Bayes estiamtes under bounded loss. Biometrika, 67, 629-638.

SMITH, J.Q. (1981). The multiparameter steady model. J.R. Statist. Soc. B, 43, 256-260.

STONE, M. (1970). Necessary and sufficient conditions for convergence in probability to invariant posterior distributions. Ann. Math. Statist. 4, 1349-1353.

STONE, J.M. (1982). Investigation of the structural properties of Kalman filter models for forecasting non-stationary time-series. Ph.D. Thesis, London University.

TONG, H. and LIM, K.S. (1980). Threshold autoregression limit cycles and cyclical data (with Discussion). J.R. Statist. Soc. B, 42, 245-292.

TRUSTRAM, K. (1971). Linear Programming. London: Routledge and Kegan Paul.

WEST, M. (1981). Robust sequential approximate Bayesian estimation. J.R. Statist. Soc. B, 43, 157-166.

WHITTLE, P. (1963). Prediction and Regulation by Linear Least-Squares. Princeton: Van Nostrand.