DYNAMIC BAYESIAN MODELS

for

VECTOR TIME SERIES ANALYSIS

&

FORECASTING

Emanuel Pimentel Barbosa

Thesis submitted for the degree of Ph.D. in Statistics

DEPARTMENT OF STATISTICS
UNIVERSITY OF WARWICK
NOVEMBER 1989

CONTENTS

CHAPTER 1 - INTRODUCTION

1.1 - Historical background 01

1.2 - Dynamic models and multivariate time series 02

1.3 - The Bayesian approach to dynamic models 02

1.4 - Implementation aspects and tractability 04

1.5 - Plan of the Thesis 04

1.6 - Terminology and notation 05

1.7 - How to read this Thesis 06

CHAPTER 2 – THE UNIVARIATE DYNAMIC LINEAR MODEL

2.1 - Model formulation and analysis 07

2.2 - Specification of the noise variances 10

2.3 - Non-informative initial priors : Reference analysis 13

2.4 - Model specification and design 16

Appendix 2.1 – Linear Bayes theory 21

Appendix 2.1a – Filtering recursive equations 23

CHAPTER 3 – UNIVARIATE EXTENSIONS TO STANDARD DLM's

3.1 - Introduction 24

3.2 - Non-linear dynamic models 25

3.3 - General non-linear problems and finite mixture of DLM's . . . 29

CHAPTER 4 – MODEL MONITORING & INTERVENTION

4.1 - Introduction 32
4.2 - Feed-forward intervention 32
4.3 - Model monitoring & feed-back intervention 34
4.4 - Monitoring schemes & its operating characteristics 37
Appendix 4.1 - Simulation results 41

CHAPTER 5 – MULTIVARIATE DYNAMIC LINEAR MODELS

5.1 - Introduction 48
5.2 - Problem formulation and general background 50
5.3 - Conjugate analysis for special DLM's 55
5.4 - Reference analysis of common component DLM's 60
Appendix 5.1 - The matrix inversion Lemma & the R.A. of DLM's 69
Appendix 5.2 - Notation and the matrix-Normal/Wishart updating 71

CHAPTER 6 – A GENERAL METHODOLOGY FOR DLM ANALYSIS

6.1 - Introduction 80
6.2 - A general multivariate DLM 81
6.3 - An alternative model representation 88
6.4 - An application with real data 96
Appendix 6.1a - Conditional independence results 99
Appendix 6.1b - Sensitivity analysis results 100
Appendix 6.2a - Generalized inverse matrix techniques 102
Appendix 6.2b - Factorization methods for p.d. matrices 103

CHAPTER 7 – SOME ALTERNATIVE TECHNIQUES FOR DLM ANALYSIS

7.1 - Introduction 113
7.2 - The Marginal approximation approach 113
7.3 - The Likelihood approximation approach 118
7.4 - The Linear Bayes approach 122
Appendix 7.1a - Matrix Taylor series expansion 125
Appendix 7.1b - The Robust filter equations 126
Appendix 7.2 - An approximated likelihood function 129
Appendix 7.3 - Linear Bayes Theory and related results 131

CHAPTER 8 – SOME SPECIAL MODELLING ASPECTS

8.1 - Introduction 146
8.2 - Modelling non-linearities 147
8.3 - Modelling compositional data 152
8.4 - Vector auto-regressive dynamic models 153
Appendix 8.1 - Updating equations for the DNLM 158
Appendix 8.2 - Covariance between quadratic forms & related results . . . 160
Appendix 8.3 - Initial prior seting for BVAR models 164

CHAPTER 9 – FURTHER MODEL EXTENSIONS AND ILLUSTRATIONS

9.1 - Introduction 165
9.2 - Special modelling extensions 165
9.3 - The SGM model : a numerical illustration 167

REFERENCES 178

ACKNOLEDGEMENTS

I would like to express my deep appreciation to Jeff Harrison for his extremely competent guidance and constant encouragement during all these years . He should be credited for many original ideas developed throughout the thesis as well as for the criticisms of all chapters of the thesis .

I am also very grateful to Mike West for many illuminating discussions during early stages of my work , and Tony O'Hagan for some important lessons and discussions about Bayesian statistics .

Also , my fellow Ph.D. students , members of the staff and visiting fellows in general should be credited for providing an excellent atmosphere for research . In particular , I would like to express my gratitude to Ezequiel Uriel and Maria Luiza Molto for their friendship , and for the provision of data about the Spanish economy .

I am also indebted to FAPESP and UFSCar , for financial support .

SUMMARY

This thesis considers the Bayesian analysis of general multivariate DLM's (Dynamic Linear Models) for vector time series forecasting where the observational variance matrices are unknown . This extends considerably some previous work based on conjugate analysis for a special sub–class of vector DLM's where all marginal univariate models follow the same structure .

The new methods developed in this thesis , are shown to have a better performance than other competing approaches to vector DLM analysis , as for instance , the one based on the Student t filter .

Practical aspects of implementation of the new methods , as well as some theoretical properties are discussed , further model extensions are considered , including non–linear models , and some applications with real and simulated data are provided .

CHAPTER 1
INTRODUCTION

1.1 - Historical background .

Although the use of conditional probability as the basis for statistical analysis can be traced back to the eighteenth century with the work of Bayes(1763) and Laplace(1774) , only more recently in this century there has been a revival of Bayesian ideas lead by De Finnetti , Jeffreys and others . Over the last fifty years there has been rapidly increasing support for the Bayesian approach to scientific learning and decision making , with the emergence of consiberable literature , as for instance , the books of Savage(1954) , Jeffreys(1961) , Lindley(1965) , DeGroot(1970) , Zellner(1971) , Box & Tiao(1973) , De Finetti(1974) , Aitchison & Dunsmore(1980) , Berger(1985) , Smith,J.Q.(1987) , O'Hagan(1988) , Press(1989) , West & Harrison(1989) and others .

On the other hand , the development of data processing methods for dealing with noise contaminated observations can be traced back to Gauss & Legendre (circa 1800) who developed , independently of each other , the method of linear least squares - Gelb(1974) . More recently , a recursive solution for linear least squares was obtained by Plackett(1950) , and Kalman and others (circa 1960) using state-space formulations designed optimal recursive filters for the estimation of (stochastic) dynamic linear systems , which has represented a significant progress in relation to the classical theory of stochastic processes based on the work of Wiener and Kolmogorov (circa 1940) .

It soon became apparent that the Bayesian approach provided a neat theoretical framework for the recursive estimation of stochastic dynamic linear systems - Ho & Lee(1964) , Aoki(1967) . In a time series & forecasting context however , at that time , these ideas of state-space models and Bayesian methods were not wide spread yet and it was time for ARIMA models - Box & Jenkins(1970) , based on classical stochastic process theory .

The Bayesian approach was developed in a time series and forecasting context by Harrison & Stevens(1971,76) with the reformulation of the state-space representation and introduction of the Dynamic Linear Model and multi-process models , leading to a methodology known as Bayesian forecasting - Harrison & West(1987) , West & Harrison(1989) , which has opened a new era for time series modelling and forecasting . This is the basis upon which we build the models and methods presented in this thesis .

1.2 - Dynamic Models and Multivariate Time Series .

In principle there are two different ways of representing a stochastic time process (time series) : the Auto-regressive / Moving-average - ARMA representation for stationary processes and the state-space or Markovian representation , and it is claimed that they are theoretically equivalent - Aoki(1987) , chapter 4 . In fact , as discussed in the next chapter , one sub-class of dynamic linear models - the constant DLM's - are equivalent to the ARIMA models . However , in practical terms there are considerable dissimilarities and the reasons in favour of the second one are as follow .

First , the state-space representation is more meaningful since the system parameters have a natural interpretation ; also they can couple with more general non-stationary data , and unlike in the computationally demanding maximun likelihood method , the estimation algorithm for the mean and variance of the process parameter are the efficient Kalman filter equations .

Second , while traditional time series analysis is primarily directed toward scalar-valued data (and usually represents the time series or its differenced version by a scalar ARMA process) , the state-space or Markovian representation treats several variables simultaneously as vector-valued variables . This allows us to understand the dynamic relations between the component series as well as to obtain more adequate forecasts since we are using more information than just one time series .

In fact , the Markovian representation of multivariate time series is more natural and simpler , not suffering the drawbacks and complexity (such as excessive number of parameters , extremely difficult model identification , etc) of a vector ARMA structure [see Tiao & Tsay(1989)] .

However , the state-space formulation on its own (without the Bayesian furnishment) as in Aoki(1987) or Harvey(1981) , does not provide the necessary and adequate environment for time series modelling and forecasting . The Bayesian forecasting approach of Harrison & West(1987) does provide this sort of environment through some facilities not present in other methods such as : probabilistic interpretation for parameters & observables , model building from simple components , intervention analysis and others .

1.3 - The Bayesian approach to Dynamic models .

The Bayesian analysis of dynamic models is essentially as neat as the usual Bayesian analysis

for static models . Let y be a set of observations (scalar or vector) and θ be a set of parameters . A typical static model is defined by the likelihood (probability distribution of the observations conditional on the parameters) $p(y/\theta)$ and the parameters' prior distribution $p(\theta)$.

By conditional probability laws (conglomerative or total probability property and Bayes' theorem) the predictive distribution p(y) and the posterior distribution $p(\theta/y)$ are obtained as :

$$p(y) = \int_{\theta} p(y/\theta).p(\theta).d\theta \qquad \& \qquad p(\theta/y) = [p(y)]^{-1}.p(y/\theta).p(\theta) \tag{1.1}$$

In a dynamic model , the process parameter θ_t , also called system state , changes as time passes according to an evolutional distribution $p(\theta_t/\theta_{t-1})$ which describes a Markovian transition from the state θ_{t-1} to θ_t . That is , given the process parameter prior distribution at time t-1 , $p(\theta_{t-1})$ and the dynamic evolution distribution $p(\theta_t/\theta_{t-1})$, the prior distribution at time t , $p(\theta_t)$ will be given by

$$p(\theta_t) = \int_{\theta_{t-1}} p(\theta_t/\theta_{t-1}).p(\theta_{t-1}).d\theta_{t-1} \tag{1.2}$$

The observations are now available or obtained sequentially in time (typically at equiespaced intervals but not necessarilly) , with the predictive distribution $p(y_t)$ and the posterior distribution $p(\theta_t/y_t)$ given respectivelly by ,

$$p(y_t) = \int_{\theta_t} p(y_t/\theta_t).p(\theta_t).d\theta_t \tag{1.3}$$

$$p(\theta_t/y_t) = [p(y_t)]^{-1}.p(y_t/\theta_t).p(\theta_t) \tag{1.4}$$

In all these equations it is implicitly assumed that these probability distributions are conditional on D_{t-1} , i.e. , on the history of the series of data up to time t-1 . Then , when we apply the system of equations (1.2)-(1.4) at time t-1 (left hand side time index) , from (1.4) we get

$$p(\theta_{t-1}/y_{t-1}) = p(\theta_{t-1}/D_{t-1}) = p(\theta_{t-1})$$

which will be input for the equation (1.2) at time t , closing the whole dynamic cycle . Clearly , these updating equations provide us a sequential learning scheme for the (unobservable) process parameter θ_t as well as a sequence of one-step-ahead predictive distributions for the observables y_t . For two-steps-ahead or more long-term forecasting or when we find some missing observations , we use (1.2)-(1.3) repeatedly but skipping (1.4) .

1.4 - Implementation aspects and tractability .

A practical implementation of equations (1.2)-(1.4) depends on the solution of the integrals present in the first two of these equations and the difficulty of this problem is related with the existence of a tractable sufficient statistic for θ_t . A sufficient statistic for a parameter θ_t summarizes all the information the observations provide about the parameters , and consequently , the distribution of the parameters conditional on the observations is the same as the distribution obtained by conditioning only on the sufficient statistic .

Apart from some special cases involving linear normal and some other models where these updating equations can be written in a neat closed form (standard conjugate prior analysis and reference analysis) , a general solution to this problem is far from trivial .

In fact , a totally general strategy to approach this problem would require the use of numerical integration procedures , which are equivalent to using a discrete approximation to the posterior distribution . Efficient N.I. procedures require initially the transformation of the parameter space to $\mathbf{R}^p$ where p is the parameter dimension , and are based typicaly on Gaussian quadrature as an interpolatory integration rule for low-dimensional integrands or on adaptive importance sampling for high-dimensional integrands - Smith,A.F.M. at all(1985) , Shaw(1987) . The implementation of efficient Gaussian quadrature methods in the context of dynamic models is presented by Pole & West(1988) , with some real applications presented in the case of dimension one .

Alternatively , an approximate Bayesian analysis for the system of equations (1.2)-(1.4) can be pursued through analytical approximations exploiting the particularities of a specific application and using adequate assumptions . Another approach for the implementation of the Bayesian paradigm is the use of Linear Bayesian estimation - Hartigan(1969) and it is also discussed in this thesis , although some of the more important results of this thesis are based on analytical approximations .

1.5 - Plan of the Thesis .

The basic univariate DLM theory and some related topics are reviewed in chapter 2 of this thesis , which is the key chapter about background material . This comprises the general model formulation , the design of each component (trend , seasonal effects , regressors , etc.) as

well as the more usual forms of analysis . Such analysis include the standard conjugate prior analysis and also the use of non-informative priors (Reference Analysis) .

It is followed by a couple of chapters about extensions of the basic univariate DLM theory . In chapter 3 we review some dynamic non-linear models with their respective analysis , which will be useful in later chapters , as well as some new ideas about non-linearity . To complete the presentation and discussion about univariate models some Bayesian monitoring procedures are reviewed in chapter 4 acompanied by a simulation study .

The main results of the thesis are presented through chapters 5 to 9 as follows . In chapter 5 we study the difficulties associated with the Bayesian analysis of a general multivariate DLM as well as the limitations of the common components multivariate DLM are showed in detail through theoretical analysis and practical examplification .

In chapter 6 a new methodology designed to overcome the restrictions of the methods of chapter 5 is presented . This includes model formulation and analysis as well as updating algorithm with full implementation details . Also some theoretical properties for the new methods are presented with the respective proofs and one example with real data is provided .

Some alternative estimation procedures for multivariate DLM analysis are presented and discussed comparatively in chapter 7 . Some results from theoretical analysis and simulation experiments are also presented .

In chapter 8 , some modelling aspects are discussed in order to extend even more the range of applications of the proposed procedures . This includes some non-linear and non-normal extensions of multivariate DLM's as well as an analysis and extension of Bayesian Vector Auto-Regressive - BVAR models .

Finally , some other modelling and related aspects are discussed in chapter 9 of this thesis , where a numerical application is provided .

1.6 - Terminology and notation .

Throughout the thesis all probability distributions are defined via densities with respect to the Lebesgue measure , and they are represented by the generic symbol p(.) .

Also , as usual in time series notation , no distinction is made between random variables and their observed values , since the context generally clarifies whether the statements relate to random variables , observations or even both .

A typical notation used throughout this thesis is the following :

$$(\underline{y}/\underline{x}) \sim N(\underline{m}, C)$$

It means that the conditional distribution of the random vector $\underline{y}$ given the value of the random vector $\underline{x}$ has a multivariate normal distribution with mean vector $\underline{m}$ and variance-covariance matrix C .

In general , vectors are underlined and matrices appear as capital letters .

Finally , it is worth remarking that the equations are numbered according to the chapters . For instance , equation (1.2) means equation number 2 of the chapter 1 . Further notation will be introduced as necessary in each chapter .

1.7 - How to read this thesis .

In principle , the material covered throughout the chapters is sequential and sometimes related to many references , but some effort has been made in order to make each chapter as independent and easy to read as possible . In this way , some background material or more technical results are presented as Appendices in most chapters .

In practice , it is suggested that , after this introductory chapter , reading starts from chapter 2 which covers some important background material and then proceeds directly to chapter 5 and the following chapters , where the main results are presented . Chapters 3 and 4 can be read as refered to in the later chapters since , although they are important reference chapters , they do not constitute the kernel of this thesis .

A second alternative would be to read the chapters in a strict sequential way .

CHAPTER 2

THE UNIVARIATE DYNAMIC LINEAR MODEL

This chapter provides background material about univariate dynamic linear models as a precursor to the study of multivariate dynamic models (chapters 5 to 9) as well as some other univariate models and related issues (chapters 3 & 4) .

The formulation of the univariate DLM as well as the simplest form of analysis for such a model is presented in section 2.1 . The problem of specification of the noise variances is discussed in section 2.2 . In section 2.3 an alternative form of analysis for the DLM is presented where non-informative in tial priors are used . Finally , in section 2.4 , the problem of model specification and design is addressed .

2.1 - Model formulation and analysis :

2.1.1 - Definition of DLM :

As we have seen in the last chapter , section 1.3 , any dynamic Bayesian model for a sequence of observations y_t (t = 1,2,...) is characterized by the following three elements : the likelihood $p(y_t / \theta_t)$, or distribution of the data given the parameters , the evolution distribution $p(\theta_t / \theta_{t-1})$ which describes the (Markovian) parameter transition from time t-1 to time t , and the initial prior distribution $p(\theta_0)$.

In this context , a general univariate (normal) Dynamic Linear Model - DLM for a sequence of observations y_t (t = 1,2,...) and parametrized by a pxl vector $\underline{\theta}_t$ is defined by :

$$\text{i) Likelihood : } (y_t / \underline{\theta}_t) \sim N(F_t^T . \underline{\theta}_t ; V_t) \tag{2.1}$$

$$\text{ii) Evolution : } (\underline{\theta}_t / \underline{\theta}_{t-1}) \sim N(G_t . \underline{\theta}_{t-1} ; W_t) \tag{2.2}$$

$$\text{iii) Initial prior : } (\underline{\theta}_0 / D_0) \sim N(\underline{m}_0 ; C_0) \tag{2.3}$$

where the quadruple $\{ F_t , G_t , V_t , W_t \}$ characterizes a specific DLM and , given $\underline{\theta}_t$, y_t is Conditionaly Independent of the past values of the series . Also , given $\underline{\theta}_{t-1}$, $\underline{\theta}_t$ is C.I. of $\underline{\theta}_{t-2}$, etc , i.e. , the parameter evolution is Markovian .

An equivalent and more usual Definition of DLM for a sequence of observations y_t (t =

1,2,...) and parametrized by a px1 vector $\underline{\theta}_t$ is given by :

$$\text{i) Observation Equation :} \quad y_t = F_t^T . \underline{\theta}_t + v_t \quad , \quad v_t \sim N(0; V_t) \tag{2.1a}$$

$$\text{ii) System Equation :} \quad \underline{\theta}_t = G_t . \underline{\theta}_{t-1} + \underline{w}_t \quad , \quad \underline{w}_t \sim N(0; W_t) \tag{2.2a}$$

$$\text{iii) Initial Information :} \quad (\underline{\theta}_0 / D_0) \sim N(\underline{m}_0 ; C_0) \tag{2.3a}$$

where the quadruple $\{ F_t , G_t , V_t , W_t \}$ is known , and the observational & evolution error sequences v_t & $\underline{w}_t$ are independent in time and of each other , and also independent of $(\underline{\theta}_0 / D_0)$.

Comments :

i) One important special case of DLM's is when $\{F_t, G_t, V_t, W_t\} = \{F, G, V, W\}$ which is called a Constant DLM and includes essentially all classical linear time series models . The reason why classical ARIMA models can be represented as constant DLM's is that any finite order difference or differential equation can be rewritten as a vector first order equation (Markovian representation) . Such equivalence is shown in detail for instance , in West(1982) , Migon(1984) , Aoki(1987) or West & Harrison(1989) . An extremelly usefull class of DLM's , which contain the class of Constant DLM's as a particular case , are the so called Time Series DLM's , where $\{ F_t , G_t \} = \{ F , G \}$ but V_t & W_t can vary in time . Other particular cases of DLM's include , for instance , static and dynamic regression models ($G_t = I$ in both cases and $\underline{w}_t = \underline{0}$ in the static case) as well as general Markovian processes ($v_t = 0$) .

ii) Apparently more general models could be defined by allowing the error sequences $\{ v_t \}$ & $\{ \underline{w}_t \}$ to be both auto-correlated and cross-correlated , and some definitions of DLM's allow for such structure , as for instance , in Ameen(1987) . However , it is always possible to rephrase such a correlated model in terms of one that satisfies the independence assumptions . Thus , we lose nothing by imposing such restrictions which lead to the simplest and most easily analysed mathematical form .

Also , the normality assumptions for the error terms v_t & $\underline{w}_t$ are not restrictive since , as shown in section 2.1.2 , the same estimated quantities are obtained with or without normality assumptions .

2.1.2 - Basic Conjugate Analysis : V known

Initially we present the analysis of the DLM supposing that V_t is known (the problem of how to specify V_t & W_t is considered in the next section) . The updating equations for the process parameter $\underline{\theta}_t$ are obtained using three different procedures : Normal theory , Bayes theorem and Linear Bayes estimation . The results are summarized as follows :

Theorem : For the univariate DLM of section 2.1.1 , one-step forecasts and posterior distributions are given , for each t , as follows :

Notation : $(\underline{\theta}_{t-1} / D_{t-1}) \sim N(\underline{m}_{t-1}, C_{t-1})$

i) Prior at time t : $(\underline{\theta}_t / D_{t-1}) \sim N(\underline{a}_t = G_t.\underline{m}_{t-1}, R_t = G_t.C_{t-1}.G_t^T + W_t)$ (2.4)

ii) 1-step forecast : $(y_t / D_{t-1}) \sim N(f_t = F_t^T.\underline{a}_t, Q_t = F_t^T.R_t.F_t + V_t)$ (2.5)

iii) Posterior at t : $(\underline{\theta}_t / D_t) \sim N(\underline{m}_t = \underline{a}_t + A_t.e_t, C_t = R_t - A_t.Q_t.A_t^T)$ (2.6)

where $A_t = R_t.F_t.Q_t^{-1}$ & $e_t = y_t - f_t$.

Proof : i) is obtained immediatelly from the system equation and ii) is a consequence , co sidering the observation equation . Using standard multivariate normal theory , the conditional distribution of $\underline{\theta}_t$ given $D_t = \{y_t , D_{t-1}\}$ is obtained and iii) follows directly .

Alternatively , iii) can be proved by Bayes Theorem , as follows :

$$p(\underline{\theta}_t / D_t) \propto p(\underline{\theta}_t / D_{t-1}).p(y_t / \underline{\theta}_t) \quad \text{where :} \tag{2.7}$$

$$p(\underline{\theta}_t / D_{t-1}) \propto exp\{-\frac{1}{2}.(\underline{\theta}_t - \underline{a}_t)^T.R_t^{-1}.(\underline{\theta}_t - \underline{a}_t)\} \quad \text{and} \tag{2.7a}$$

$$p(y_t / \underline{\theta}_t) \propto exp\{-\frac{1}{2.V_t}.(y_t - F_t^T.\underline{\theta}_t)^2\} \tag{2.7b}$$

After some algebra , we get the following posterior ,

$$p(\underline{\theta}_t / D_t) \propto exp\{-\frac{1}{2}.(\underline{\theta}_t - \underline{m}_t)^T.C_t^{-1}.(\underline{\theta}_t - \underline{m}_t)\} \tag{2.8}$$

with $\underline{m}_t$ & C_t as given in iii) above .

Comment 1 (Linear Bayes Estimator) : One interesting aspect of the updating equations for the posterior moments of the DLM is that they can be derived using Linear Bayes methods without the normality assumption . In fact , from Appendix 2.1 , under a quadratic loss function , the linear Bayes estimate M of a random vector X given the data Y and its associated expected squared error C are given respectively by ,

$$M = E(X) + A.(Y - EY) \tag{2.9}$$

$$C = V(X) - A.V(Y).A^T \tag{2.10}$$

where $A = Cov(X, Y).[V(Y)]^{-1}$.

In the notation of DLM's , we have $X = \underline{\theta}_t$ with prior moments $EX = \underline{a}_t$ & $V(X) = R_t$, and $Y = y_t$ with moments $EY = f_t$ & $V(y_t) = Q_t$. As a consequence ,

$$A = Cov(X, Y).[V(Y)]^{-1} = Cov(\underline{\theta}_t, y_t).[V(y_t)]^{-1} = R_t.F_t.Q_t^{-1}$$

and we get , using the LBE of equations (2.9)-(2.10) , the same updating equations given by equation (2.6) .

Comment 2 (Forecast Functions) : The forecast function $f_t(k)$ is defined for all integers $k \geq 0$ as

$$f_t(k) = E\{\mu_{t+k}/D_t\} = E\{F'_{t+k}.\underline{\theta}_{t+k}/D_t\}$$

where $\mu_{t+k} = F'_{t+k}.\underline{\theta}_{t+k}$ is the mean response function . For k strictly greater than 0 , the forecast function provides the expected values of future observations given current information , $f_t(k) = E\{y_{t+k} \ D_t\}$. In the special but very important case of Time Series DLM's , where $\{F_t, G_t\} = \{F, G\}$ for all t , the forecast function is given by $f_t(k) = F'.G^k.\underline{m}_t$. The form of the forecast function in k is a major guide to the design of DLM's (specification of F and G) and this will be considered briefly in section 2.4 of this chapter .

Comment 3 (Filtering & Smoothing) : The use of current data to revise inferences about previous values of the process parameter is called filtering . This an important tool for retrospective time series analysis where the information recently obtained is filtered back to previous time points . The distribution of $(\underline{\theta}_{t-k} / D_t)$ for $k \geq 1$ and any fixed t , is called the k-step filtered distribution for the state vector at that time . In this context , a related concept is that of smoothing a time series . The retrospective estimation of a time series mean response function μ_t using the filtered distributions (μ_{t-k} / D_t) for $k \geq 1$ is called smoothing the series . At any given time t , such filtered distributions are derived recursively backwards in time , as shown in Appendix 2.1a

2.2 - Specification of the Noise Variances

2.2.1 - Observational variance learning : In most practical situations , it is not realistic to expect that we know with accuracy the values of V_t . More likely , the kind of information available will be beliefs about certain features of this sequence , as for instance , that V_t

is constant but unknown . In such cases , a fully conjugate Bayesian learning procedure is available as detailed below .

Assuming a normal-inverse gamma prior distribution for $(\underline{\theta}_t , V)$, with n degrees of freedom and mean S^{-1} (S is the shape parameter) ,

$$(\underline{\theta}_t , V / D_{t-1}) \quad \sim \quad N\Gamma^{-1}_{n_{t-1}} (\underline{a}_t , R_t , S_{t-1}) \tag{2.11}$$

where $\underline{a}_t$ & R_t are obtained as in section 2.1.2 , the joint posterior distribution for these parameters will be given by ,

$$(\underline{\theta}_t , V / D_t) \quad \sim \quad N\Gamma^{-1}_{n_t} (\underline{m}_t , C_t , S_t) \quad \text{where :} \tag{2.12}$$

$$\underline{m}_t = \underline{a}_t + A_t . (y_t - F_t^T . \underline{a}_t) \tag{2.13}$$

$$A_t = R_t . F_t . Q_t^{-1} \tag{2.13a}$$

$$Q_t = F_t^T . R_t . F_t + S_{t-1} \tag{2.13b}$$

$$C_t = (R_t - A_t . Q_t . A_t^T) . S_t / S_{t-1} \tag{2.14}$$

$$S_t = S_{t-1} . (n_{t-1} + e_t^2 / Q_t) / n_t \tag{2.15}$$

$$n_t = n_{t-1} + 1 \tag{2.15a}$$

The marginal posterior distribution for $\underline{\theta}_t$ will be a multivariate t distribution with parameters n_t , $\underline{m}_t$ & C_t and the marginal posterior distribution for V will be a gamma (chi-square) distribution with parameters S_t & n_t .

Note that for a large number of d.f. , V will be approximately equal to S_{t-1} , equal with probability one as n_{t-1} tends to infinity , when the t distribution approach normality . In the limit therefore , these updating equations are equivalent to those derived in section 2.1.2 . In fact , the basic difference between the case where we know V and the present case , is the presence of a kind of correction factor S_t / S_{t-1} in the equation (2.14) that will tend to the unit when the d.f. increase and then , this equation will approach equation (2.6) .

The results of this section are presented in more detail and with many references in West & Harrison(1989) . Also , the more general methods presented in chapter 6 of this thesis , for multivariate DLM analysis , are shown to coincide with the results of this section , in the particular case of dimension one (scalar models) .

Comment : Sometimes , when the observation errors $\{v_t\}$ are not normally distributed with constant variance , we may need to use some previous transformation in the data (for instance , power transformations , as in Box & Cox(1964)) in order to restore constancy in time of observational variances . However , such procedures stop effective intervention and for this reason we recommend the use of variance laws instead of variance stabilization . For discussion and references about variance laws in a time series context , see West & Harrison(1989) , chapter 10 .

When it is not possible or easy to identify such systematic changes of variance in time , or when there is some extra stochastic variation in the observations , we can model this variation in V , increasing its uncertainty from time t-1 to t through the use of a discount factor δ , which correspond to rewriting the equations (2 15)-(2.15a) as follows :

$$S_t = S_{t-1}.(\delta.n_{t-1} + e_t^2 \; Q_t) \; n_t \qquad (2.15')$$

$$n_t = \delta.n_{t-1} + 1 \qquad (2.15a')$$

where in practice δ takes large values , typicaly like .98 or .99 .

2.2.2 - Specification of W_t : The discount method

The specification of a suitable structure for the system noise variance matrices W_t is crucial in the implementation of the DLM updating equations . The elements of W_t quantify the increase of uncertainty or loss of information about $\underline{\theta}_t$ from time t-1 to time t . Concretely , from section 2.1.2 (equation 2.4) , we have ,

$$V(\underline{\theta}_t \,/\, D_{t-1}) = R_t = G_t.V(\underline{\theta}_{t-1} \,/\, D_{t-1}).G_t^T + W_t = P_t + W_t$$

and it is natural to think in terms of a rate of decay of information δ such that $R_t = P_t \,/\, \delta$ for some scalar discount factor δ $(0 < \delta \leq 1)$. Since $W_t = P_t.(1-\delta)/\delta$, this implies an increase in variance or loss of information about $\underline{\theta}_t$ from time t-1 to time t , of $100(1-\delta)/\delta\,\%$. This procedure , proposed originally by Ameen & Harrison(1985) , has a strong intuitive appealing and has overcome most of the difficulties in specifying W_t . In fact , these discount factors play a role analogous to those used in non-Bayesian point forecasting methods , in particular to exponential smoothing techniques - Abraham & Ledholter (1983) , providing interpretation and meaning within the DLM framework .

However , the theoretical equivalence between a discount DLM and a standard one is not always guaranteed , since $W_t = R_t - P_t$ specified in this way fails sometimes to be a proper variance matrix . In order to avoid such problems , the discount method can be modified by discounting each diagonal block of P_t and not the whole matrix as before . As suggested by Harrison & West(1987) , each block of the G matrix (the model is formed by the superposition of blocks , as discussed in section 2.4) with corresponding blocks in C_{t-1} should be discounted with a constant discount factor , with perhaps different factors for each block .

Although W_t can be specified by other methods , as for instance in Gamerman(1987) , because of the reasons mentioned before as well as our own experience in using the discount method , we will consider throughout this thesis , the implementation of DLM's via blocks discounting .

2.3 - Non-informative initial Priors : Reference analysis

2.3.1 - Introduction

We present here an alternative analysis for the DLM where it is not necessary to specify the hyper-parameters (such as $\underline{m}_0$, C_0 , etc) in the initial priors as in the analyses of section 2.1.2 or 2.2.1 . This is done through the use of vague or non-informative prior distributions (reference priors) as the ones proposed by Jeffreys(1961) , and constitutes a particular case of the more general results presented in chapter 5 of this thesis concerning multivariate models .

The so called Jeffreys' rule for multiparameter problems [see Box & Tiao (1973) , pg 54] can be stated as follows : ' The initial prior distribution for a set of parameters is taken to be proportional to the square root of the determinant of the information matrix '.

In the concrete case of a prior distribution for $(\underline{\theta}_t , V)$ we shall first of all assume that $\underline{\theta}_t$ and V are approximately independent . Then , considering the standard (locally uniform) reference prior for $\underline{\theta}_t$, given by $p(\underline{\theta}_t) \propto constant$, we have as joint reference prior , $p(\underline{\theta}_t, V) \propto p(V)$. Considering the Jeffreys' rule for multiparameter problems , we have

$$p(\underline{\theta}_t , V) \propto |I(V)|^{\frac{1}{2}}$$

where $I(V)$ is the information matrix , i.e. , minus the expected value of the second derivative of the log-likelihood . Since , $p(\underline{\theta}_t, V^{-1}) = p(\underline{\theta}_t, V).\frac{\partial V}{\partial V^{-1}}$, we have , differentiating the

log-likelihood twice with respect to V^{-1} , the following result

$$I(V) = I(V^{-1}).[\frac{\partial V}{\partial V^{-1}}]^{-2} \propto [\frac{\partial V}{\partial V^{-1}}]^{-1} = V^{-2}$$

Consequently , the reference form will be given by ,

$$p(\underline{\theta}_t , V) \propto V^{-1}$$

2.3.2 - Reference Analysis of DLM's : Theory

Theorem : For the univariate (normal) DLM defined by the equations (2.1a)-(2.2a) , let the initial prior information be represented by the reference form $p(\underline{\theta}_t , V) \propto V^{-1}$. Then , assuming that W_t has full rank , we have :

i) The joint prior and posterior distributions for $(\underline{\theta}_t , V)$ at time t = 1,2,... are given by

$$p(\underline{\theta}_t, V/D_{t-1}) \propto V^{-(1+\frac{a_t-1}{2})}.exp\{-\frac{1}{2V}(\underline{\theta}_t^T.H_t.\underline{\theta}_t - 2.\underline{\theta}_t^T.T_t + L_t)\}$$

$$p(\underline{\theta}_t, V/D_t) \propto V^{-(1+\frac{a_t}{2})}.exp\{\frac{1}{2V}(\underline{\theta}_t^T.K_t.\underline{\theta}_t - 2.\underline{\theta}_t^T.U_t + E_t)\} \quad \text{where :}$$

$$H_t = W_t^{-1} - W_t^{-1}.G_t.Z_t^{-1}.G_t.W_t^{-1} \tag{2.16}$$

$$Z_t = G_t^T.W_t^{-1}.G_t + K_{t-1} \tag{2.16a}$$

$$T_t = W_t^{-1}.G_t.Z_t^{-1}.U_{t-1} \tag{2.17}$$

$$K_t = H_t + F_t.F_t^T \tag{2.18}$$

$$U_t = T_t + F_t.y_t \tag{2.19}$$

$$L_t = E_{t-1} - U_{t-1}^T.Z_t^{-1}.U_{t-1} \tag{2.20}$$

$$E_t = L_t + y_t^2 \tag{2.21}$$

$$a_t = a_{t-1} + 1 \tag{2.22}$$

with $H_1 = 0, T_1 = 0 \quad L_1 = 0, a_1 = 0$ as initial setup .

ii) For $t \geq t_P$, where $t_P = p + 1$, the posterior distribution for $(\underline{\theta}_t, V/D_t)$ is a normal-inverse chi square distribution with parametrization given by

$$(\underline{m}_t = K_t^{-1}.U_t , C_t = K_t^{-1} , S_t = E_t - U_t.\underline{m}_t , n_t = a_t - 1) \tag{2.23}$$

Proof : The proof is by finite induction on t and the details can be found in Pole & West(1987) . In fact , this theorem is a particular case of a more general theorem concerning multivariate DLM's presented in section 5.4 of this thesis with the respective proof .

2.3.3 - Reference Analysis of DLM's : Implementation

Although the updating equations defined in the previous theorem , part (i) , are valid for all $t \geq 0$, for computational and interpretation reasons (avoidance of matrix inversions , easier interpretation , etc) , it is preferable to use the standard updating equations (section 2.2.1) for $t \geq t_P$, since both algorithms are algebraicly equivalent at these time points . For $0 \leq t < t_P$ however , where there is no such equivalence , we need to implement the reference analysis algorithm , and one major difficulty is the setup of the covariance matrices W_t . Since the posterior covariance matrices C_t do not exist for $t < t_P$ (K_t does not have full rank) we can not apply the traditional discount techniques used in the implementation of standard conjugate analysis of DLM's . The procedure used here is in line with Pole & West(1987) , and consists in assuming $W_t = 0$ for t = 1,2,., t_P . This practical procedure has its rationale in the fact that is not possible to detect or estimate any changes in parameters during the first t_P observations , since we have only one observation for each parameter in $\underline{\theta}_t$ or V , and so , we lose nothing by setting them to zero . The basic result necessary for the practical implementation of these ideas is given by the following theorem .

Theorem : In the context of Reference Analysis for the univariate (normal) DLM , suppose that G_t is non-singular and $W_t = 0$. Then , the prior and posterior distributions of $\underline{\theta}_t$ and V have the same forms as in the theorem of section 2.3.2 , with the same initialization and observation updating equations but different time updating equations , as follows :

$$H_t = G_t^{T^{-1}}.K_{t-1}.G_t^{-1} \tag{2.24}$$

$$T_t = G_t^{T^{-1}}.U_{t-1} \tag{2.25}$$

$$L_t = E_{t-1} \tag{2.26}$$

Proof : Since G_t is non-singular , the system equation can be inverted , giving $\underline{\theta}_{t-1} = G_t^{-1}.\underline{\theta}_t$, which is a linear transformation with constant Jacobian . Then , supposing that the joint posterior distribution at time t-1 has the stated form as in the last theorem , the joint prior distribution is obtained as ,

$$p(\underline{\theta}_t, V/D_{t-1}) \propto V^{1+\frac{n_{t-1}}{2}}.exp\{-\frac{1}{2V}(\underline{\theta}_t^T.H_t.\underline{\theta}_t - 2.\underline{\theta}_t^T.T_t + L_t)\}$$

where $H_t = G_t^{T^{-1}}.K_{t-1}.G_t^{-1}$; $T_t = G_t^{T^{-1}}.U_{t-1}$ & $L_t = E_{t-1}$. Since the joint prior distribution and the likelihood are the same considered in the last theorem , the joint posterior distribution will be the same and the theorem is proved by induction .

Corollary : As a consequence of the last two theorems of section 2.3 , a practical Reference Analysis algorithm for the univariate (normal) DLM is given as follows :

i) For $t = 0, 1, .., t_P - 1$, where $t_P = p + 1$ is the minimum time such that the posterior distributions are proper , use the updating equations of the last theorem .

ii) For $t \geq t_P$, use the standard normal-inverse chi square updating algorithm of section 2.2.1 .

2.4 - Model Specification & Design

2.4.1 - Basic concepts : Observability & Similarity

Observability is a fundamental concept in linear systems theory and its counterpart in a time series DLM context , is related with the identifiability of the p-dimensional state vector $\underline{\theta}_t$ from the knowledge of the mean response parameter over time $(\mu_t, \mu_{t+1}, ...)$.

In order to introduce this concept , we consider initially the case where $W_t = 0$ for all t , so that $\underline{\theta}_t = G.\underline{\theta}_{t-1}$ and $\mu_{t+k} = F'.G^k.\underline{\theta}_t$. Clearly , at least p distinct values of the mean response are required for such identification , with parametric parsimony suggesting that no more than p be necessary . The p distinct values starting at t , denoted by $\underline{\mu}_t = (\mu_t, \mu_{t+1}, .., \mu_{t+p-1})'$ are related to the state vector $\underline{\theta}_t$ via $\underline{\mu}_t = T.\underline{\theta}_t$ where ,

$$T = \begin{pmatrix} F' \\ F'.G \\ \vdots \\ F'.G^{p-1} \end{pmatrix} \tag{2.27}$$

is the observability matrix . Thus , to determine the state vector $\underline{\theta}_t$ from $\underline{\mu}_t$ precisely we require that T be non-singular and then $\underline{\theta}_t = T^{-1}.\underline{\mu}_t$.

These ideas of parsimony and identifiability of parameters in the case of purely deterministic evolution $(W_t = 0)$ motivate the formal definition of observability in the general case :

Any TSDLM $\{F, G, ., .\}$ is observable if and only if the pxp observability matrix T given by equation (2.27) has full rank p .

The concept of observability allows a modeller to restrict attention to a sub-class of DLM's that are parsimoniously parametrized but provide the full range of forecast functions . This sub-class is still large and any given form of forecast function may typically be derived from many observable models ; such group or class of observable models with the same forecast function form are called Similar Models . Since the forecast function pattern is related with

the eigen-structure of the system matrix , two observable TSDLM's are similar if and only if the corresponding system matrices have identical eigenvalues .

An equivalent definition is that two observable models $M_1 = \{F_1, G_1\}$ & $M_2 = \{F_2, M_2\}$ are Similar if they are related to each other through a linear transformation , i.e. , if there exist a non-singular matrix H (similarity matrix) such that $F_1' = F_2'.H^{-1}$ & $G_1 = H.G_2.H^{-1}$. In this way , the similarity of observable models is defined via the similarity of their respective system matrices .

2.4.2 - System Eigen-Structure & Canonical Models

Within each group of similar models , we identify particular models with specific , simple structure that provide Canonical DLM's consistent with the required forecast function . Since the forecast function form is related to the eigen-structure of the system matrix G , we consider here the various possible configurations that may , and do arise in practice . The two basic kinds of configurations are the case of real eigenvalues and the case of complex eigenvalues for G , presented as follows :

Case I : G has a single real eigenvalue λ of multiplicity p .

There are clearly an infinite number of such matrices , the simplest one being $G = \lambda.I$, which is not an useful representation because it implies a non-observable model . The class of observable models with G as stated above for this case , is restricted to those whose system matrix is similar to the Jordan block $J_p(\lambda)$ given by ,

$$J_p(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & .. & 0 \\ 0 & \lambda & 1 & & : \\ 0 & 0 & \lambda & & 0 \\ : & & & \ddots & 1 \\ 0 & .. & .. & 0 & \lambda \end{pmatrix} \qquad (2.28)$$

Also , by observability constraint , it can be shown that F must have its first element non-zero , and the simplest observable form is given by $E_p' = (1,0,...,0)$. Then , in practice , for this case , the simplest canonical form is ,

$$\{ F , G \} = \{ E_p' , J_p(\lambda) \} \qquad (2.29)$$

As a consequence , the corresponding forecast function for this model block will be given by ,

$$f_t(k) = E_p'.[J_p(\lambda)]^k.\underline{m}_t = \lambda^k.\sum_{j=0}^{p-1} a_{jt}.k^j \qquad (2.30)$$

which is a general polynomial form multiplied by the factor λ^k , where the coefficients a_{jt} do not depend on k but only on λ and $\underline{m_t}$. In the important special case of $p = 2$, we have ,

$$f_t(k) = (1\ 0).\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}^k . \begin{pmatrix} m_{1t} \\ m_{2t} \end{pmatrix} = (m_{1t} + \frac{k}{\lambda}.m_{2t}).\lambda^k \qquad (2.30a)$$

For $\lambda = 1$, we have the linear forecast function $f_t(k) = m_{1t} + k.m_{2t}$, which constitutes one of the most useful trend representations . This model , characterized by $F' = (1\ 0)$ & $G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ may be viewed as representing a locally linear development of the mean response function over time and it is called sometimes a linear growth model .

Case I'- multiple real eigenvalues : G has s distinct eigenvalues $\lambda_1, .., \lambda_s$ with λ_i having multiplicity $p_i \geq 1$ so that $p = p_1 + ... + p_s$. In this case , it follows that G is similar to the block diagonal Jordan form matrix ,

$$J = diag\{J_{p_1}(\lambda_1), ..., J_{p_s}(\lambda_s)\} \qquad (2.31)$$

defined by the superposition of Jordan blocks $J_{p_i}(\lambda_i)$, (i = 1,..,s) , one for each of the distinct eigenvalues λ_i and having dimension given by the corresponding multiplicities p_i . Also , F is similar to $E' = \{E'_{p_1}, ..., E'_{p_s}\}$ and in practice , for this case I', we specify G = J and F = E , with J and E as defined above .

In this case , it is clear that the general forecast function will be given by

$$f_t(k) = \sum_{i=1}^{s} f_{it}(k)$$

where , for each i , $f_{it}(k)$ has its form given by (2.30) .

Case II : G has a pair of complex conjugate eigenvalues $\lambda_1 = \lambda.e^{iw}$ & $\lambda_2 = \lambda.e^{-iw}$ for some real λ and ω .

In principle , this model is similar to any model of the form $\{(1,1)', diag(\lambda_1, \lambda_2), ., .\}$ but in practice we do not use this canonical model since this would imply a complex parametrization . Instead , we identify a real canonical form for G using the transformation matrix $H = \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}$, which gives $E_2 = (1\ 1)'.H^{-1} = (1\ 0)'$ and ,

$$J_2(\lambda, w) = H.\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.H^{-1} = \lambda.\begin{pmatrix} cos\, w & sin\, w \\ -sin\, w & cos\, w \end{pmatrix} \qquad (2.32)$$

and in practice we specify $F = E_2$ and $G = J_2(\lambda, w)$.

As a consequence , the forecast function for this model will be given by

$$f_t(k) = F'.G^k.\underline{m}_t = E_2'.[J_2(\lambda,w)]^k.\underline{m}_t = E_2'.J_2(\lambda^k,kw).\underline{m}_t$$
$$= (m_{1t}.cos\,kw + m_{2t}.sin\,kw).\lambda^k = \lambda^k.A_t.cos\,(kw+\phi_t) \qquad (2.33)$$

where $A_t = m_{1t}^2 + m_{2t}^2$ is the amplitude , and $\phi_t = arc\,tan\,(-\frac{m_{2t}}{m_{1t}})$ is the phase-angle , or just , phase of the periodic (also called harmonic) model component . The quantity ω is the frequency of the periodic component , and defines the number of time intervals over which the harmonic completes a full cycle , this number being $2\pi/w$. It is interesting to notice that for $\omega = \pi$ (Nyquist frequency) , we have just one real eigenvalue equal to $-\lambda$ (note that $e^{i\pi} = e^{-i\pi} = cos\,\pi = -1$) and not a pair of complex eigenvalues .

The forecast function thus has a sinusoidal form , that is modified by the multiplicative term λ^k determined by λ , which will dampen , explode or keep unchanged ($\lambda < 1$, $|\lambda| > 1$, $\lambda = 1$) the periodic component . This last case ($\lambda = 1$) leads to a pure cosine wave of period $2\pi/w$ and is often used to model seasonal time series .

Case II' : G has a set of distinct pairs $\lambda_j.e^{\pm i.w_j}$, (j = 1,..,k) , for a given integer k and given pairs $\{\lambda_j, w_j\}$ of reals . In this case , the real canonical representation for the elements F & G will be given by the superposition of the elements given in the case II above . Then , we have the canonical observable representation

$$F = (1,0,1,0,...,1,0) \quad \& \quad G = diag\{J_2(\lambda_1,w_1),...,J_2(\lambda_k,w_k)\} \qquad (2.34)$$

Also , the forecast function for this case will be given by the sum of k terms like (2.33) .

Finally , we should mention that multiplicities associated with each pair of conjugate complex eigenvalues have not been discussed because such models are not common in practice . However , further details are to be found in West & Harrison(1989) .

2.4.3 - Standard Component Models and The Superposition Principle

In practice , by standard component models we mean canonical TSDLM's that represent the most common block models such as Trend component (linear or otherwise) , seasonal components (harmonic blocks with different frequencies) or regression components . In this way , according to the developments of the last sub-section , a typical linear trend block will

be represented by $\left\{\begin{pmatrix}1\\0\end{pmatrix}' ; \begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}\right\}$, a typical harmonic block for seasonal data in the frequency ω will be represented by $\left\{\begin{pmatrix}1\\0\end{pmatrix}' ; \begin{pmatrix} cos\, w & sin\, w\\ -sin\, w & cos\, w\end{pmatrix}\right\}$, and so on .

More complex models where an unknown eigenvalue λ (in the previous two examples , the eigenvalue was equal to 1) can appear in the system matrix G constitute a case of non-linear model and will be considered only in the next chapter .

It is worthwhile mentioning that models can be represented in a variety of observable forms using any bijective reparametrisation . In particular , polynomial trends and seasonal components for DLM's can be also represented in alternative ways , other than the ones considered in this thesis , based on Jordan blocks . For example , a DLM with polynomial trend of degree p-1 can be built by taking $F_t' = (1,0,..,0)$ and $G_t = G$ an upper triangular matrix such that the non-zero triangle is the Pascal triangle , i.e. , $G_{i,j} = \binom{j-1}{i-1}(i \leq j)$ (i,j = 1,..,p) where $(i \leq j)$ means 1 if $i \leq j$ and zero otherwise . In fact , both representations of polynomials - the Pascal representation based on powers of k , $\{1, k, .., k^{p-1}\}$ and the Jordan representation based on standard factorial polynomials $\left\{\binom{k}{0}, \binom{k}{1}, .., \binom{k}{p-1}\right\}$ or even an alternative representation where G is an upper triangular matrix formed only by 1's - give not only the same values for their forecast functions , but the representations themselves coincide in the linear case , i.e. , $G = \begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$ in the Jordan or Pascal or triangular 1's representation .

Also , the seasonal components can be represented both by Fourier form , using trigonometric functions , as developed in the last sub-section (the complex eigenvalue case) or alternatively , using the so called form free seasonal effects - West & Harrison(1989) , chapter 8 , where the system matrix is represented as a permutation matrix . The reason why we consider the Fourier form of representation for seasonal phenomenon in this thesis is mainly parsimony and orthogonality , and both trend and seasonal components are represented by Jordan blocks .

Another model component of practical importance is a regression component $\{ F_t = x_t ; G_t = 1 \}$ designed to represent the effect of a regressor variable x_t in the model . Such variables are typicaly external regressors or lagged versions of external or response variables (in the last case we have an auto-regressive component model) .

Then , these three types of modelling components - Trend , Seasonal & Regression component - constitute the basic blocks that builds up a widely applicable set of models . For instance , we may find out that a given real time series of monthly observations can be represented by

a linear trend block plus four harmonic components (with 12 , 6 , 4 & 3 months periodicity) and an external regressor variable . In this case , each one of these 6 blocks or components are DLM's and they should be combined together to form the whole model , which should be also a DLM .

The general , simple and extremely useful principle that guarantees that a linear combination or superposition of independent DLM's is itself a DLM is called the Superposition Principle . This enables the construction of complex models by adding simple block structures . Formally , one version of this principle can be stated as follows :

The Superposition Principle : For integer $k > 1$, consider the k time series y_{jt} generated by the DLM's $\{F_j, G_j, V_j, W_j\}$ with state vectors $\underline{\theta}_{jt}$ of dimensions p_j (j = 1,..,k) . Assume that , for all distinct i & j $(1 < j, i < k)$ the error series v_{jt} & $\underline{w}_{jt}$ are mutually independent of the series v_{it} & $\underline{w}_{it}$. Then , the series defined by $y_t = \sum_{j=1}^{k} y_{jt}$ follows a DLM $\{F, G, V, W\}_t$ with state vector given by $\underline{\theta}_t' = (\theta_{1t}', .., \theta_{kt}')$ of dimension $p = \sum p_j$ and such that $F_t' = \{F_{1t}', .., F_{kt}'\}$ & $G_t = diag\{G_{1t}, .., G_{kt}\}$. Also , $V_t = \sum V_{jt}$, $W_t = diag\{W_{1t}, .., W_{kt}\}$ and the forecast function of y_t is given by the superposition of the forecast functions of the k components .

Proof : It is immediate from linear theory , and in the case of normal DLM's , from Normal theory .

Finally , it is important to mention that the independence of the component models in the superposition principle is not crucial and can be relaxed in order to provide a more wider statement . In fact , a more general statement of superposition assumes joint normality for both observation and system error series , but not necessarily independence .

Appendix 2.1 : Linear Bayes Theory

Some Basic Concepts :

Lets f(Y) be the Bayes Estimator for a random variable X . Then , the Bayes risk is defined by ,

$$r(f) = E\{f(Y) - X\}^2 = E_Y E_{X/Y}\{f(Y) - X\}^2 = E_Y d(f/Y)$$

where d(f/Y) is the posterior expected squared error , i.e., the posterior expected loss for a quadratic loss function . Clearly , r(f) is minimised completely by minimizing d(f/Y) for each Y , which gives f(Y) = E(X/Y) .

However , this is an arbitrarily complicated function of Y and requires full Bayesian analysis to derive . Linear estimation simplifies the derivation by restricting f(Y) to the linear class $f(Y) = a + b^T.Y$. Then , a and b are chosen to minimise r(f) , which gives the optimal estimate

$$\hat{f}_X(Y) = \hat{a} + \hat{b}^T.Y$$

where the coeficients $\hat{a}$ and $\hat{b}$ are given respectively by

$$\hat{a} = E(X) - Cov(X,Y).[Var(Y)]^{-1}.E(Y)$$
$$\hat{b} = Cov(X,Y).[Var(Y)]^{-1}$$

Also , $V_X = r(\hat{f})$ is a measure of how good $\hat{f}(Y)$ is . It is the expected squared error , but notice that it is the unconditional expectation , i.e. , the prior expectation of posterior expected squared error .

Now , suppose X is a vector . In general , we could estimate each component X_i using a linear function of a different predictor vector Y_i given by $f_i(X_i) = a_i + b_i^T.Y_i$. Then , a_i and b_i will be chosen to minimize $r_i(f_i) = E\{f_i(Y) - X_i\}^2$ and a measure of accuracy is provided by $V_{X_i} = r_i(\hat{f}_i)$, what gives a kind of expected posterior variance for $\hat{f}_i$.

We would also like an analogue of covariance , and this is obtained as follows . Let ,

$$f(Y) = \begin{pmatrix} f_1(Y_1) \\ \vdots \\ f_k(Y_k) \end{pmatrix} \qquad \& \qquad \hat{f}(Y) = \begin{pmatrix} \hat{f}_1(Y_1) \\ \vdots \\ \hat{f}_k(Y_k) \end{pmatrix}$$

and define $r(f) = E\{f(Y) - X\}\{f(Y) - X\}^T$ and $V_X = r(\hat{f})$.

The diagonal elements of V_X will be V_{X_i} and the off-diagonal elements will be the expected cross-products of errors , which will correspond to covariances .

Then , the (optimal) linear Bayes estimate of the random vector X (given the data vector Y) and its associated expected squared error are given respectively by:

$$\hat{f}_X(Y) = E(X) + A.(Y - E(Y))$$
$$V_X = Var(X) - A.Var(Y).A^T$$

where $A = Cov(X,Y).[Var(Y)]^{-1}$.

Appendix 2.1a - Filtering Recursive Equations

Theorem : In the univariate DLM $\{ F_t, G_t, V_t, W_t \}$, considering the usual notation adopted in this chapter , define $B_t = C_t.G'_{t+1}.R^{-1}_{t+1}$ for all t . Then , for all k such that $1 \leq k \leq t$, the filtered distributions for the process parameter are given by $(\underline{\theta}_{t-k} / D_t) \sim N\{\underline{a}_t(-k); R_t(-k)\}$ where ,

$$\underline{a}_t(-k) = \underline{m}_{t-k} + B_{t-k}.[\underline{a}_t(-k+1) - \underline{a}_{t-k+1}]$$
$$R_t(-k) = C_{t-k} - B_{t-k}.[R_{t-k+1} - R_t(-k+1)].B'_{t-k}$$

with starting values given by $\underline{a}_t(0) = \underline{m}_t$ & $R_t(0) = C_t$. Also , $\underline{a}_{t-k+1} = \underline{a}_{t-k}(1)$ & $R_{t-k+1} = R_{t-k}(1)$.

Proof : The filtered densities are defined recursively via

$$p(\underline{\theta}_{t-k} / D_t) = \int p(\underline{\theta}_{t-k}/\underline{\theta}_{t-k+1}, D_t).p(\underline{\theta}_{t-k+1}/D_t).d\underline{\theta}_{t-k+1}$$

which suggests proof by induction on k . For details , see for instance , West & Harrison(1989) , chapter 4 .

Consequences : i) If $V_t = V = \phi^{-1}$ is unknown and the conjugate analysis of section 2.1 applied , then ,

$$(\underline{\theta}_{t-k}/D_t) \sim T_{n_t}[\underline{a}_t(-k); (S_t/S_{t-k}).R_t(-k)]$$

ii) The corresponding smoothed distributions for the mean response of the series are given by

$$(\mu_{t-k}/D_t) \sim T_{n_t}[f_t(-k); (S_t/S_{t-k}).F'_{t-k}.R_t(-k).F_{t-k}]$$

where , in an extension of the notation for the forecast function to negative arguments $f_t(-k) = F'_{t-k}.\underline{a}_t(-k)$.

CHAPTER 3

UNIVARIATE EXTENSIONS TO STANDARD DLM's

3.1 - Introduction

As we have seen in the last chapter (section 2.4) , one of the most common types of components in a DLM structure include cases of models where the system matrix G has eigenvalues with modulus equal to one . That is the case for instance , of polynomial trends , standard regressions and harmonic components for seasonal data .

However , important models are found in practice where there is one or more unknown eigenvalues in the system matrix (G is not totally known as before) , and this implies that the mean response function $\mu_t = f(\underline{\theta}_t)$ cannot be represented as a linear function of $\underline{\theta}_t$ as in the standard DLM . Dynamic models in which the mean response function has non-linear terms in some parameters will be called generically non-linear models .

The most common kind of non-linear models found use in practice in a time series modelling and forecasting context , are such that the mean response is a bilinear function of the state vector (bilinear models) , which opens even further the range of applications for dynamic models . Some very useful modelling structures such as non-linear (Gompertz type) growth models , linear growth models with multiplicative seasonality , transfer response models (special regression models with lagged variables) and ARMA components for noise representation , are all typical examples of non-linear or bilinear extensions of standard DLM's .

Since these models present extra unknown quantities that break the neat linear formulation , the analysis will not be as simple as before (chapter 2) and some analytical approximation or numerical integration will be required . In the next section of this chapter we present briefly some examples of bilinear models as well as a general formulation and analysis based on Taylor series expansion , which will be extended to the multivariate case in chapter 8 of this thesis . Also in this section the concept of non-linearity is discussed in association with the geometric concept of curvature , which permits us to introduce the class of 'close-to-linear models' as a very important subset of non-linear models . In fact , the examples that will be presented , belong to this special category of non-linear models , and that is why the analytical approximations based on Taylor series expansions produce such good results in practice .

A totaly general strategy to the analysis of non-linear dynamic models however demands the use of numerical integration techniques , which is discussed in section 3 of this chapter and

discussed further in chapter 9 of this thesis , in a multivariate DLM context , where the extra unknown quantities are not elements of G , but elements of V .

3.2 - Non-linear dynamic models

3.2.1 - Examples of bilinear dynamic models

Example 1 : Seasonal growth multiplicative models

In many time series where seasonality is a major factor , it happens that amplitudes of seasonal components increase significantly at higher levels of the series . And , one typical situation found in practice is characterized by multiplicative seasonal variation which is proportional to the process level .

In this case , the seasonal effects component ρ_t and the level (trend) component γ_t are linked together to form the (bilinear) mean response function $\mu_t = \gamma_t.(1+\rho_t)$ where $\gamma_t = F_1'.\underline{\theta}_{1t}$ and $\rho_t = F_2'.\underline{\theta}_{2t}$ are linear functions of the process parameter $\underline{\theta}_t = \begin{pmatrix} \underline{\theta}_{1t} \\ \underline{\theta}_{2t} \end{pmatrix}$ and F_1 & F_2 are known vectors of 0 and 1's .

The full model of course is defined as usual by the observation equation $y_t = \mu_t + v_t$ and the system equation $\underline{\theta}_t = G.\underline{\theta}_{t-1} + \underline{w}_t$ where μ_t is a non-linear function of $\underline{\theta}_t$ as defined before , and the F's & G blocks for the linear trend and seasonal components are represented in canonical forms as presented in chapter 2 , section 2.4 .

For more discussion , references or application of this model using real data , see for instance , Migon(1984) or West & Harrison(1989) .

The full presentation and analysis of an extension of this model to the multivariate case is provided in chapter 8 of this thesis .

Example 2 : Transfer Function Models

In a standard DLM framework , consider regression on current and lagged values of a single independent or input variable X_t (up to a maximum lag k , say) . Assuming initially that the regression coefficients are constant over time , the mean response μ_t will be given by ,

$$\mu_t = \beta_0.X_t + \beta_1.X_{t-1} + ... + \beta_k.X_{t-k}$$

This represents a kind of 'form-free' transfer function model , where μ_t represents the effect of the input variable X (since time t-k to time t) on the response variable Y .

Although very flexible , this model structure can be non-parsimonious and inappropriate for cases in which it is felt that the lagged effects persist into the future , perhaps decaying towards zero as time progress . One simple way of adapting the regression structure to incorporate such features , is to consider regression not on X_t directly , but on a constructed effect variable ξ_t measuring the combined effect of current and past values of X , as follows :

$$\text{(i) observation equation : } y_t = \xi_t + v_t$$

$$\text{(ii) system equation : } \xi_t = \lambda.\xi_{t-1} + \rho.X_t + \omega_t$$

where $0 \leq \lambda \leq 1$ and $\rho \geq 0$. This is obviously a non-linear model , since it can be rewritten with $\underline{\theta}_t = \begin{pmatrix} \xi_t \\ \rho \end{pmatrix}$ and $G_t = \begin{pmatrix} \lambda & X_t \\ 0 & 1 \end{pmatrix}$ where λ is unknown .

For more details about these models , as well as applications and references , see for instance , West & Harrison(1989) . For multivariate extensions of transfer function models , see chapter 8 of this thesis , where the form free model is discussed in a vectorial context .

Example 3 : Bayesian Auto-Regressive models

In conection with the last Example 2 , it is important to mention that there is an alternative way to make the 'form-free' transfer function a more parsimonious model , even when the lagged effects persist into the future and k is large . This is possible , introducing 'stochastic constraints' in the $\beta's$ parameters through the use of convenient priors for that parameters .

One special but very important case of such modelling structure occurs when the inputs are lagged values of the response variable , that is , when both the response and the input variable (in a transfer function model) coincide . In this case , we have the so called Bayesian auto-regressive model , where the use of convenient priors for the auto-regressive parameters is a key feature in order to obtain model parsimony (k can be large , if necessary) .

In chapter 8 of this thesis we present in detail the multivariate counterpart of these auto-regressive models , the so called BVAR (Bayesian Vector Auto-Regressive) models , which can be seen as a special multivariate DLM .

3.2.2 - Approximated analysis of non-linear dynamic models

From the examples and concepts discussed early in this chapter , it is clear that many models with parameter non-linearities may be written as

$$i) \text{ observation equation : } \quad y_t = F_t(\underline{\theta}_t) + v_t$$

$$ii) \text{ system equation : } \quad \underline{\theta}_t = G_t(\underline{\theta}_{t-1}) + \underline{w}_t$$

where $F_t(.)$ is a given non-linear regression function mapping the p vector $\underline{\theta}_t$ to the real mean response , $G_t(.)$ is a given non-linear vector evolution function , and v_t & $\underline{w}_t$ are error terms subject to the usual assumptions .

One of the simplest and most easily interpretable approaches to non-linear models is based on the use of Taylor series expansions to the mean response and/or evolution function . Briefly , using our standard DLM notation , the application of Taylor series linearization around the expected value $\underline{a}_t$ of $\underline{\theta}_t$ in the non-linear regression function , gives us

$$F_t(\underline{\theta}_t) = F_t(\underline{a}_t) + F_t^*.(\underline{\theta}_t - \underline{a}_t) + \text{higher order terms}$$

where $F_t^* = \left\{\frac{\partial F_t(\underline{\theta}_t)}{\partial \underline{\theta}_t}\right\}_{\underline{\theta}_t = \underline{a}_t}$, and in practice the higher order terms are neglected . Using similar Taylor series expansion for the evolution equation , we obtain the following 'linearized' DLM ,

$$y_t = g_t + F_t^*.\underline{\theta}_t + v_t$$

$$\underline{\theta}_t = \underline{h}_t + G_t^*.\underline{\theta}_{t-1} + \underline{w}_t$$

where $g_t = F_t(\underline{a}_t) - F_t^*.\underline{a}_t$; $\underline{h}_t = G_t(\underline{m}_{t-1}) - G_t^*.\underline{m}_{t-1}$ & $G_t^* = \left\{\frac{\partial G_t(\underline{\theta}_{t-1})}{\partial \underline{\theta}_{t-1}}\right\}_{\underline{\theta}_{t-1} = \underline{m}_{t-1}}$. In practice , the extra terms g_t & $\underline{h}_t$ do not bring any extra difficulty into the model analysis since both terms are known .

One obvious refinement of the simple linearization method is to consider the inclusion of quadratic terms in the Taylor series expansions . Considering the non-linear regression function $F_t(\underline{\theta}_t)$ expanded till second order terms , we have

$$\mu_t = F_t(\underline{\theta}_t) = F_t(\underline{a}_t) + F_t^*.(\underline{\theta}_t - \underline{a}_t) + \frac{1}{2}.(\underline{\theta}_t - \underline{a}_t)'.T_t.(\underline{\theta}_t - \underline{a}_t)$$

where $T_t = \left\{\frac{\partial^2 F_t(\underline{\theta}_t)}{\partial \underline{\theta}_t \partial \underline{\theta}_t'}\right\}_{\underline{\theta}_t = \underline{a}_t}$ is a kind of curvature matrix .

In order to carry on the analysis , it will be necessary to evaluate quantities such as $E(\mu_t/D_{t-1})$ & $Var(\mu_t/D_{t-1})$, which will require expressions for the first two moments of a random quadratic form . Based on normality assumptions , after some algebra , we get the following expressions ,

$$E(\mu_t/D_{t-1}) = F_t(\underline{a}_t) + \frac{1}{2}.tr\, T_t.R_t$$

$$Var(\mu_t/D_{t-1}) = F_t^{*'}.R_t.F_t^* + \frac{1}{2}.tr\,\{T_t.R_t\}^2$$

where $R_t = Var(\underline{\theta}_t / D_{t-1})$.

The first result is easily obtainable (see for instance , Seber(1977)) . The second one is a bit more laborious and constitutes a particular case of a more general result presented in the Appendix 8.1c , chapter 8 of this thesis . It is also shown there , concerning the seasonal growth multiplicative model of example 1 , that after some algebra , the expressions for both moments ($E(\mu_t / D_{t-1})$ & $Var(\mu_t / D_{t-1})$) coincide with the results for that model given by Migon(1984) .

Comments :

i) It is important to observe that , although the linearized model has a simple form of analysis as in the standard DLM case , its mean response function keeps the basic non-linear structure as in the original formulation .

ii) The truncation of the Taylor series expansion or neglect of higher order terms , is a key assumption and its justification is based on smoothness and well behaved properties of the non-linear functions .

3.2.3 - Approximation assessment and non-linearity measures

i) As mentioned before , most of the more important (non-linear) models found in practice , belong to the special class of bilinear (or second order) models . Such bilinear models , as for instance the seasonal growth multiplicative model , present only second order nonlinearities , and only the first two derivatives of the mean response function are non-zero .As a consequence , we are not neglecting any non-zero term in the Taylor series expansion of $\mu_t = F_t(\underline{\theta}_t)$ and , no truncation approximation is involved .

In fact , the expression for $E(\mu_t / D_{t-1})$ that we are using is exact and does not depend even on normality . For the variance expression as given above however , an approximation is involved , since we are evaluating variances of quadratic terms . In other words , we are calculating 4^{th} order moments , but only the first two moments are considered in the model . Thus , such 4^{th} order moments should be related to the first two moments under some probability distribution assumption , and we use normality assumptions in order to get practical results .

ii) Another important aspect to be mentioned is in connection with the quality of the approximations involved and is related to the extent of ' non-linearity' itself . The approximations

involved became better and better when the combination model & data present only a small amount of non-linearity , that is , when the model curvature is small . Each regression model has associated with it a surface called a solution locus - Box & Lucas(1959) , and non-linearity at any point can be quantified through the curvature of this surface at the given point .

Such measures of non-linearity for regression models , extensivelly studied by Bates & Watts(1980) and others , depend basically on the second derivatives of the regression function in relation to the model parameters . Although the context of dynamic models is not exactly the same as static regression models , it is clear that such measures are associated with the curvature matrix $T_t = \left\{ \frac{\partial^2 F_t(\underline{\theta}_t)}{\partial \underline{\theta}_t \, \partial \underline{\theta}_t'} \right\}_{\underline{\theta}_t = \underline{a}_t}$, which is a null matrix for linear models .

In particular , when the second term in the variance expression $Var(\mu_t / D_{t-1})$ is small compared with the first one , the curvature will be small and the model will be called a ' close-to-linear model ' .

3.3 - General 'non-linear' problems and finite mixture of DLM's

3.3.1 - Introduction

Although very useful in practice for the analysis of many important non-linear models , the analytical approximation methods presented in the last section of this chapter are not totally general and cannot cope with all sorts of unknown elements in a DLM framework . For instance , if we have unknown elements not in G but in one of the variance matrices (V or W) , we need more general techniques . For this reason , since in later chapters of this thesis we will need to deal with unknown elements in the matrix V in a multivariate DLM context , we devote some attention to these more general techniques , also useful for usual non-linear problems . We consider initially the non-linear problem as an illustration .

3.3.2 - Non-linear models and finite mixture of DLM's

One basic fact about the class of non-linear models we have been considering in this chapter is that conditionally on given values , for a certain sub-set of the unknown parameters , the models became linear (conditionally linear models) .

For easy of notation , consider that there is just one unknown eigenvalue λ in the G matrix responsible for the non-linearity in the mean response function . For any given value for λ , the standard DLM analysis applies , and our problem is how to learn about this parameter

from the data . With λ constant over time though unknown , the formal Bayesian analysis proceeds as follows :

(i) For each value $0 < \lambda < 1$, specify an initial prior distribution for $\underline{\theta}_0$ (and also for V , if unknown) $p(\underline{\theta}_0 / \lambda, D_0)$ and also $p(\lambda / D_0)$.

(ii) For each value of λ , process the data according to the usual sequential updating equations for the DLM , with $G = G(\lambda)$, which gives , at time t , the predictive density $p(y_t / \lambda, D_{t-1})$ and the posterior $p(\underline{\theta}_t / \lambda, D_t)$.

(iii) For each value of λ , obtain the posterior for λ via Bayes'theorem ,

$$p(\lambda / D_t) = c \, . \, p(\lambda / D_{t-1}) \, . \, p(y_t / \lambda, D_{t-1})$$

where the integral $c^{-1} = \int_0^1 p(y_t/\lambda, D_{t-1}).p(\lambda/D_{t-1}).d\lambda$ should be evaluated , analytically or numerically .

(iv) Posterior inferences for $\underline{\theta}_t$ and other quantities of interest should be evaluated , as for instance ,

$$p(\underline{\theta}_t / D_t) = \int_0^1 p(\underline{\theta}_t / \lambda, D_t).p(\lambda / D_t).d\lambda$$

Since , in general there is no tractable or easily calculable expression for the last couple of integrals , in order to obtain an explicit solution to this problem , we adopt here a very pragmatic and simple approach .

Consider a finite and fixed (time invariant) discretization of the parameter space for λ , namely , $\{ \lambda_1 , \lambda_2 , ..., \lambda_k \}$ for some integer $k \geq 1$. Then , the learning process for λ will be as follows :

i) Consider $p_{t-1}(j) = p(\lambda_j / D_{t-1}) = Pr\{\lambda = \lambda_j / D_{t-1}\}$ for all t , with specified initial prior probabilities $p_0(j)$, (j = 1,2,..,k) .

ii) The likelihood function for λ will be given , at time t , by $l_t(j) = p(y_t / \lambda_j, D_{t-1})$ (j = 1,2,...,k) , from the usual DLM updating equations .

iii) The posterior probabilities for λ_j are updated , as usual , by Bayes'theorem , giving $p_t(j) = c_t . p_{t-1}(j) . l_t(j)$, (j = 1,2,...,k) where $c_t^{-1} = \sum_{j=1}^{k} p_{t-1}(j) . l_t(j)$.

iv) The unconditional posterior distribution for $\underline{\theta}_t$ or any other unconditional distribution is obtained as a finite mixture of the k conditional distributions , using $p_t(j)$ as the corresponding weights , as for instance ,

$$p_t(\underline{\theta}_t / D_t) = \sum_{j=1}^{k} p(\underline{\theta}_t / \lambda_j, D_t) . p_t(j)$$

As a consequence , the unconditional posterior moments will be given by similar finite mixtures of conditional moments , as follows

$$\underline{m}_t = E(\underline{\theta}_t / D_t) = \sum_{j=1}^{k} p_t(j).E(\underline{\theta}_t/\lambda_j, D_t)$$

$$C_t = V(\underline{\theta}_t / D_t) = \sum_{j=1}^{k} p_t(j).\{V(\underline{\theta}_t/\lambda_j, D_t) + \underline{m}^*_{jt}.\underline{m}^{*'}_{jt}\}$$

where $\underline{m}^*_{jt} = E(\underline{\theta}_t/\lambda_j, D_t) - \underline{m}_t$. This whole process defined by (i)-(iv) above is called multiprocess class I model - Harrison & Stevens(1976) .

Comments :

i) Such mixtures of standard models are quite widely used in non-linear models or similar problems and a good reference is West & Harrison(1989) , chapter 12 . In the engineering & control literature , analogues of mixtures of normal DLM's (each with known variances) were used , for instance , by Sorenson & Alspach(1971,72) , under the denomination of Gaussian sums . For more references to this topic (sometimes under the heading of parallel processing) , see Anderson & Moore(1980) , chapters 9 & 10 .

ii) The basic point we should make clear here is that we have approximated integrals by simple sums . This particular discretization process , in fact , is the simplest possible technique of numerical integration , equivalent say , to a kind of rectangular rule where the grid of points $\{\lambda_1, \lambda_2, .., \lambda_k\}$ is arbitrary and fixed for all time . Obviously , the accuracy of the approximation increases with the number of grid points k , but the computational demands can be enormous for a large k . Thus , the use of more refined techniques of numerical integration can be necessary . A discussion about the use of more efficient NI strategies for Bayesian analysis based on Gaussian quadrature - Smith et al(1985,87) , in a DLM context , is found in West & Harrison(1989) , chapter 13 .

CHAPTER 4

MODEL MONITORING & INTERVENTION

4.1 - Introduction

In this chapter we discuss one important characteristic of Bayesian forecasting models related with the fact that they are open to intervention whenever this is judged necessary . By model intervention we mean ammendments to the probability distribution of the process parameters anticipating major changes in the process (feed-forward intervention) or detecting and correcting performance deterioration (feed-back intervention) . In particular , we discuss in this chapter the monitoring of model performance via Bayes'factors , in the context of univariate models , which is extended to the multivariate case in chapter 9 of this thesis .

One of the simplest situations related to feed-forward intervention is when there is information concerning a future discrepant observation (outlier) as for instance , reflecting the effects of a coming strike on production levels of an industrial good or the effects of extremely bad weather on agricultural production , etc . If such outliers are expected to occur in the near future , an obvious intervention procedure is simply to ignore such wild observation or , to associate a very large variance to it . In practice , a general procedure for introducing feed-forward intervention will require not only variance ammendments but changes to the both mean and variance of the state vector .

On the other hand , when no information is available for anticipatory intervention , and the model performance starts to deteriorate , it is still possible to intervene after the detection of model inadequacies signalled by a monitoring scheme . Such monitoring schemes are based on sequences of cumulative Bayes'factors as an assessment tool for possible model inadequacies and constitutes the basic ingredient for the implementation of feed-back intervention .

In the next section we discuss briefly the more important aspects related with feed-forward intervention . In section 4.3 we discuss the basic ideas related to monitoring schemes and feed-back intervention . Finally , in section 4.4 , some aspects of Bayesian monitoring in connection with more classical statistical procedures are discussed and , in particular , the random variable 'Run Length' is studied in this context .

4.2 - Feed-Forward Intervention

The general form of feed-forward intervention in a Bayesian dynamic model and in particular in the DLM , is simply to change the prior moments of $\underline{\theta}_t$, previously $(\underline{a}_t , R_t)$, to new values $(\underline{a}_t^* , R_t^*)$, anticipating changes in the series .

Such strategy , based on these new prior moments , is adequate for forecasting further into the future , but a problem arises when considering filtering and smoothing the series using retrospective analysis . It is clear that the new post–intervention prior distribution $(\underline{\theta}_t / I_t, D_{t-1}) \sim N(\underline{a}_t^* , R_t^*)$ is no longer consistent with the DLM structure previous to time t .

In order to have such desirable consistency , we initialy observe that the previous moments for the state vector and their corresponding post–intervention values (with the stars'notation) can be related through a linear transformation , as follows :

(Lemma) - Let $\underline{\theta}_t^* = K_t.\underline{\theta}_t + \underline{h}_t$ where $\underline{\theta}_t$ has moments $(\underline{a}_t, R_t)$, K_t is a p square upper-triangular , non-singular matrix , and $\underline{h}_t$ is a p vector . Then , $\underline{\theta}_t^*$ has moments $(\underline{a}_t^*, R_t^*)$ if K_t and $\underline{h}_t$ are chosen as $K_t = U_t.Z_t^{-1}$ & $\underline{h}_t = \underline{a}_t^* - K_t.\underline{a}_t$, where U_t & Z_t are the unique , upper triangular , non-singular square root matrices of R_t^* & R_t respectively , thus $R_t^* = U_t.U_t^{'}$ & $R_t = Z_t.Z_t^{'}$.

Proof : The matrices U_t & Z_t exist and are unique since R_t^* and R_t are symmetric positive definite matrices (see , for example , Graybill,1969) . They define the Cholesky decomposition of these variance matrices and are easily computed . From the definition of $\underline{\theta}_t^*$ it follows that $\underline{a}_t^* = K_t.\underline{a}_t + \underline{h}_t$, and so , the expression above for $\underline{h}_t$ is obtained for any given K_t . Also , $R_t^* = K_t.R_t.K_t^{'}$, and thus , $U_t.U_t^{'} = (K_t.Z_t).(K_t.Z_t)^{'}$. Now , $K_t.Z_t$ is a square , non-singular , upper triangular matrix and , since the matrix U_t is unique , it follows that $U_t = K_t.Z_t$. The expression for K_t follows since Z_t is non-singular .

Consequence (Theorem) : Suppose that the moments $(\underline{a}_t^*, R_t^*)$ are specified to incorporate feed-forward intervention in a DLM and K_t & $\underline{h}_t$ are defined as in the previous Lemma . Then , the post–intervention prior is the prior obtained in a standard DLM with evolution equation amended to $\underline{\theta}_t = G_t^*.\underline{\theta}_{t-1} + \underline{w}_t^*$, $\underline{w}_t^* \sim N(\underline{h}_t, W_t^*)$ where $G_t^* = K_t.G_t$ & $W_t^* = K_t W_t K_t^{'}$.

Proof : It is immediate from the previous Lemma .

Thus , any interventions modelled by $(\underline{\theta}_t / I_t, D_{t-1}) \sim N(\underline{a}_t^*, R_t^*)$ can be formally and routinely incorporated into the model by appropriately amending the evolution equation at

time t , and reverting to the usual equations for future times not subject to intervention . It means that the usual updating , forecasting , filtering and smoothing algorithms apply directly with the post-intervention model . For applications or further details , see West & Harrison(1989) .

4.3 - Model Monitoring & feed-back intervention

4.3 1 - Bayes' Factors

This section discusses the use of automatic methods for the sequential monitoring of forecast perf rmance based on a statistical measure of accuracy . Central to such ideas is the notion of assessing model performance relative to that obtained through using one or more alternative models and the key concept is that of the Bayes' factor or ratio of two predictive distributions .

Consider any two models (denoted by M_0 and M_1) with the same structure , differing only through the values of defining parameters such as , for instance , mean and/or variance M_0 is the routine or standard model in use , and M_1 is an alternative model providing assessment of M by comparison . At any time t , each model provides a predictive distribution for y_t given D_{t-1} , given respectively by $p_0(y_t/D_{t-1})$ & $p_1(y_t/D_{t-1})$.

The Bayes'Factor for M_0 versus M_1 based on the observed value y_t is defined as $H_t = p_0(y_t/D_{t-1})/p_1(y_t/D_{t-1})$. For integers k = 1,2,..,t , the (cumulative) Bayes'factor for M_0 versus M_1 based on the sequence of k consecutive observations $y_t, y_{t-1}, .., y_{t-k+1}$ is defined as

$$H_t\ k) = \prod_{j=t-k+1}^{t} H_j = p_0(y_t, y_{t-1}, .., y_{t-k+1}/D_{t-k})\ p_1(y_t, y_{t-1}, .., y_{t-k+1}/D_{t-k})$$

These Bayes'factors , alternatively called Weights of Evidence [Jeffreys (1961) , Good(1985)] , provide the basic measures of predictive performance of M_0 relative to M_1 – see also Zellner,A.(1978) . For each k , $H_t(k)$ measures the evidence provided by the most recent k consecutive observations . Some basic features of Bayes'factors are noted :

(i) For $k = 1$, we have $H_t(1) = H_t$ for all t . Also , taking $k = t$, the Bayes'factor based on all the data is $H_t(t)$

(ii) Evidence for or against the model M_0 accumulates multiplicatively as data is processed , that is , for each $t > 1$, we have $H_t(k) = H_t.H_{t-1}(k-1)$ (k = 2,..,t) . Alternatively , on

the log scale , evidence is additive , with

$$log\ H_t(k) = log\ H_t + log\ H_{t-1}(k-1)\ ,\ k = 2,..,t$$

(iii) Following Jeffreys(1961) , a log Bayes'factor of 1 (-1) indicates evidence in favour of model M_0 (M_1) , a value of 2 or more (-2 or less) indicating the evidence to be strong . Clearly , the value 0 indicates no evidence either way .

4.3.2 - Model monitoring

In a monitoring context the focus is on local model performance and the key elements are both the individual measures H_t and the cumulative measures $H_t(k)$ for $k < t$. The cumulative measures are necessary to detect small or gradual changes not detectable individualy , but the individual measures are also necessary to detect sudden changes otherwise masked from previous evidence in favour of the standard model .

In fact , we want to know the most likely point of structural change in a given time series by identifying the most discrepant group of recent consecutive observations . This is done by minimising the Bayes ' factors $H_t(k)$ with respect to k , as follows . Let $L_t = \min_{1 \le k \le t} H_t(k)$ for t = 1,2,.. with $L_1 = H_1$, which can be written sequentially as $L_t = H_t . min\{1, L_{t-1}\}$ for t = 2,3,.. . The minimun at time t is taken at $k = l_t$, with $L_t = H_t(l_t)$ where the integers l_t , called run length , are sequentialy updated via

$$l_t = \begin{cases} l_{t-1} + 1 \ , & \text{if } L_{t-1} < 1 \\ 1 \ , & \text{if } L_{t-1} \ge 1 \end{cases}$$

Note that $L_t = H_t$ if and only if $l_t = 1$, otherwise $L_t = H_t . L_{t-1}$ and $l_t = l_{t-1} + 1$, providing the stated results .

The sequence $\{ L_t \}$ provides a sequential monitor or tracking of the predictive performance of M_0 relative to M_1 . In particular , if at time t-1 , the evidence favours M_0 so that $L_{t-1} \ge 1$, then $L_t = H_t$ and decisions about possible model inadequacies are based on y_t alone . If H_t is very small then y_t is a possible outlier or may indicate the onset of change .

Specifically , let τ ($0 < \tau < 1$) be a prespecified threshold for Bayes'factors , defining the lower limit on acceptability of L_t (in practice , typical values for τ are chosen between 0.1 and 0.2) .

One key ingredient of a monitoring scheme that should be specified is the alternative model M_1 . In fact , both predictive (one-step ahead forecasting) distributions for M_0 and M_1 , in

a DLM context , will be normal or student t distributions and , without loss of generality , the standard model is considered with mean zero and variance 1 . Perhaps the simplest example of structure for the alternative model would be to keep the same probability distribution as in M_0 but differing by a shift in the level (mean) parameter . One interesting case of modelling shifts in the level process is when we consider the possibility of level change symmetrically in both directions , in the alternative model . In this case , we have two alternative models and the associated double monitoring scheme is essentially similar to standard backward CUSUM techniques - Harrison & Davies(1964) .

A more general alternative model however that can cope with changes both in mean and variance is the scale shift model M_1 in which $e_t = y_t$ has standard deviation k rather than unity , giving , in the normal case , the Bayes ' factor $H_t = k.exp\{-0.5.e_t^2.(1-k^{-2})\}$. These models are particularly useful in modelling outliers , as for instance , in Box & Tiao(1968) , or Smith & Pettit(1985) .

4.3.3 - Model adaptation and feed-back intervention

The monitoring of model performance to detect deterioration in predictions needs to be supplemented with techniques for diagnosis of the problem and subsequent adaptation to restore predictive performance . In accordance with West & Harrison(1989) , the following logical scheme provides a guide to the use of Bayes' factors in detecting and diagnosing model breakdown . At time t , proceed with the monitor as follows :

(i) Calculate the Bayes' factor H_t . If $H_t \geq \tau$, then y_t is viewed as consistent with M_0 and we proceed to (ii) to assess the possibility of model failure prior to time t . Alternatively , if $H_t < \tau$, then y_t is a potential outlier and should be treated as a missing value for updating purposes . However , the possibility that y_t presages changes in model parameters must be catered for after rejecting the observation , thus the need for intervention is signalled and we proceed to (iii) .

(ii) Calculate the cumulative Bayes' factor L_t and the corresponding run-length l_t to assess the possibility of changes prior to time t . If $L_t \geq \tau$, then M_0 is satisfactory and so proceed to (iv) to perform standard updates , etc . Otherwise , $L_t < \tau$ indicates change that should be signalled , requiring intervention ; then , proceed to (iii) .

(iii) Issue signal of possible changes consistent with deterioration of predictions from M_0

and call for feed-back interventions to adapt the model for the future . In practice , such interventions can be implemented very effectivelly through the use of a more heavily discounted version of the matrix W_t (use of a smaller discount factor ρ , instead of the value considered for standard operation) . Following such interventions , update the time index to t+1 for the next observation stage , and proceed to (i) , reinitialising monitoring by setting $l_t = L_t = 0$.

(iv) Perform usual analysis and updating with M_0 , proceeding to (i) at time t+1 .

In practice , we have one operational problem of choosing appropriate values for τ and ρ (respectively the threshold and the discount parameter) in order to implement the above monitoring scheme . This and related issues are discussed in the next section of this chapter .

4.4 - Monitoring schemes & its operating characteristics

4.4.1 - Introduction

The monitoring of real observable processes in order to detect possible structural changes is an important practical problem with a wide range of applications in forecasting (as we have seen in the previous section of this chapter) and process control . It consists in using a scheme of sequential hypothesis tests in order to decide at each time , based on process observations , if there is or not some kind of structural change in the process . One concrete situation typical in quality control of industrial processes is to test sequentially for possible level or variability change in observable manufacturing characteristics .

Obviously , any proposed sequential test should have its Operating Characteristics well known . One of the most important of such characteristics is the number of observations taken before the detection of change - a measure of quickness of change detection - known in the literature as Run Length , the stopping rule of the process . The mean of this random variable , the Expected Run Length - ERL , and the probability of false detection of change , known as False Alarm Probability , have a role in sequential test theory similar to the traditional Type I and Type II errors in classical fixed sample size tests .

In a non-Bayesian context , Page(1954) proposed sampling inspection schemes based on Cumulative Sum of observations to test level change in i.i.d. binomial processes , Kemp(1958,71) studied the operating characteristics of such tests and Barnard(1959) proposed practical graphical representations for such tests . One good review of the whole subject is De Bruyn(1968) . For applications of such tests in a forecasting context , one pioneering work is Harrison &

Davies(1964) , and in a more sophisticated multiprocess modelling context , Ameen & Harrison(1985) .

The study of such operating characteristics , namely the ERL of CUSUM tests , can be approached in different ways , and the more common are : Simulation , as in Bissel(1969) and others , Solution of Integral Equations , as in Goel & Wu(1971) for instance , and Markov Chains , as in Brook & Evans(1972) . Obviously , this methodology applies to Bayesian monitoring schemes such as the one introduced in the previous section , since Cumulative Bayes Factors can be written in a logarithm scale as a Bayesian CUSUM .

The operating characteristics , namely the ERL and the probability of false alarm , are studied via simulation , in the next sub-section , where two different kinds of structural change are considered : level change and variability change .

4.4.2 - Simulation Study

In order to study the performance of the Bayesian monitoring scheme (section 4.3) in detecting level changes , we have simulated 1000 sequences of i.i.d. normal data with unitary variance and different values for the level θ . Each sequence finishes when the change is detected and this simulation process is repeated for different values of the operating parameters r and ρ . The average of this thousand run lengths obtained in each configuration is then calculated .

We have considered seven different values for the process level θ (0 0.5 1.0 1.5 2 3 5) with zero level for M_0 and a 6x7 matrix of values for the operating parameters (r from 0.2 to 0.5 , and ρ from 0.05 to 0.3) . The test procedure is defined by the comparison of $Z_t = ln\, H_t + min\,[0\,,\, Z_{t-1}]$ with r , where $ln\, H_t = -0.5.[ln\,\rho + (1-\rho).y_t^2\,]$ is the log-Bayes factor .

The case $\theta = 0$, corresponds to the false alarm case . The other cases represent small level changes (θ = 0.5 or 1) , medium level changes ($\theta = 1.5$) or large level change ($\theta = 2$, 3 or 5) . All the simulation cases were run in a PC-OLIVETTI M-24 , and some of the results are given at the end of this chapter .

The simulation results show that the monitoring scheme under study is reasonably robust for the ERL with respect to the operating parameters in the case of large level changes . For instance , a level change of 5 s.d. is detected in 1 step , and a level change of 3 s.d. is detected

, depending very little on the values of τ and ρ , at an expected time between 1.2 and 1.4 approximately . However , the time of detection of a (small) change of 1 s.d. can change from approximately 5 time units ($\rho = 0.3$ and $\tau = 0.5$) to approximately 15 ($\rho = 0.05$ and $\tau = 0.2$) .

In all cases , the ERL decreases with the increase of ρ and τ , but the false alarm probabilities increases , which makes the choice of these parameters a non-trivial problem . The ERL results are an informative guide for choosing the threshold and the discount values , and all depends on the relative importance of quick detection and risk of false detection . In particular , in an industrial context for instance , if we know approximately the costs associated with each situation (false detection and delay in detecting a true change) , as in Goel & Wu(1973) , this choice can be made by considering a specific loss function .

For some of the mentioned configurations , we have obtained the sample run length distribution , which for small or medium shifts shows a distribution approximately exponential . For large level changes , only the tail of the distribution has an exponential shape , and this is in accordance with some theoretical results found in Kemp(1971) and others . These graphics are shown in the end of this chapter for the case of $\theta = 1.5$.

Also , in a similar simulation study for variability change , we observe that the way the ERL changes with τ and ρ for a given value of θ is similar to the previous cases , with the exception that the detection of change is slower .

Finally , we conclude this simulation study , with one simple application of Bayesian monitoring in a control context , where the data is not i.i.d. as in a usual forecasting context , but correlated . In the control of nuclear material , known in the literature as Nuclear Material Accounting - NMA , the basic interest is to detect possible loss of nuclear material by monitoring sequences such as , M.U.F = I(n-1) - I(n) + T(n) , where I(n) and I(n-1) are respectively the accounting of nuclear material at the end and beginning of period n ; T(n) is the transfer (inputs minus outputs) of nuclear material during the period n , and M.U.F. is the material unaccounted for , which should be zero if there was no loss or error of any kind .

A detailed description of this problem is found for instance , in Goldman at al(1982) . Speed & Chulpin(1986) sugest for these M.U.F. sequences a simple model considering only first-order auto-correlations , and Downing at al(1978) consider for these sequences a MA(1) model .

Because of the difficulty in obtaining real data , we have considered the simulation of a

MA(1) process . Using this simulated data , it was calculated the ERL or average time before loss detection , for different configurations as shown in Tab-3 at the end of this chapter . One of the main aspects of these results is that the ERL values do not change very much in comparison with the i.i.d. case .

APPENDIX 4.1 -- SIMULATION RESULTS

TAB 1A : ARL - NORMAL CASE (1000 simulations / small level change)

LEVEL = 0.5

σ \ ρ	.05	.10	.15	.20	.25	.30
.2	43.13	33.06	28.83	27.32	27.32	27.16
.25	35.01	26.65	24.06	23.09	22.70	22.52
.3	27.95	22.82	20.28	19.05	19.53	19.53
.35	25.73	20.27	17.44	16.65	15.74	16.54
.4	21.59	17.02	14.59	13.96	13.53	13.74
.45	20.07	15.08	13.88	12.70	12.14	11.96
.5	17.31	13.47	12.00	10.99	10.56	10.52

LEVEL = 1.0

σ \ ρ	.05	.10	.15	.20	.25	.30
.2	14.85	11.17	10.96	10.00	9.68	9.46
.25	12.75	10.16	8.82	8.75	8.48	8.69
.3	11.09	8.78	7.83	8.00	7.61	7.68
.35	10.66	8.42	7.07	7.48	6.74	6.79
.4	9.43	7.66	7.08	6.27	6.11	5.93
.45	8.26	6.41	5.86	5.69	5.35	5.26
.5	7.56	5.95	5.61	4.92	4.98	5.15

Cont. (medium level change & false alarm case)

LEVEL = 1.5

.	.05	.10	.15	.20	.25	.30
.2	6.11	4.86	4.55	4.37	4.37	4.36
.25	5.50	4.66	4.31	4.02	4.00	3.89
.3	4.81	4.23	3.66	3.63	3.74	3.61
.35	4.43	3.89	3.85	3.59	3.35	3.40
.4	4.14	3.52	3.40	3.27	3.04	3.11
.45	4.26	3.43	3.18	3.04	2.93	2.93
.5	3.78	3.23	3.07	2.83	2.76	2.80

FALSE ALARM CASE (LEVEL = 0)

.	.05	.10	.15	.20	.25	.30
.2	89.99	68.08	58.86	55.77	56.79	55.82
.25	69.28	52.07	45.66	43.18	41.91	41.55
.3	56.43	42.45	36.42	34.80	33.84	33.73
.35	44.99	33.42	29.59	27.31	27.29	27.03
.4	39.27	29.07	25.01	22.91	22.91	21.84
.45	34.94	25.75	22.48	20.37	20.11	19.97
.5	30.41	21.94	18.56	18.27	16.32	16.34

TAB 1B : ARL - NORMAL CASE (1000 simulations / large level change)

LEVEL = 2

.	.05	.10	.15	.20	.25	.30
.2	3.11	2.81	2.62	2.52	2.46	2.55
.25	2.98	2.47	2.45	2.39	2.34	2.41
.3	2.71	2.35	2.29	2.16	2.16	2.20
.35	2.48	2.21	2.16	1.98	2.04	2.04
.4	2.44	2.23	2.04	2.01	1.94	1.99
.45	2.25	2.12	1.97	1.89	1.88	1.87
.5	2.21	1.94	1.94	1.81	1.77	1.76

LEVEL = 3

.	.05	.10	.15	.20	.25	.30
.2	1.46	1.35	1.36	1.34	1.40	1.38
.25	1.38	1.34	1.33	1.31	1.31	1.32
.3	1.36	1.28	1.27	1.27	1.26	1.26
.35	1.33	1.26	1.25	1.23	1.21	1.23
.4	1.25	1.24	1.19	1.21	1.22	1.19
.45	1.24	1.20	1.19	1.19	1.18	1.18
.5	1.23	1.18	1.17	1.16	1.15	1.15

TAB 1C : VARIABILITY CHANGE - 1000 simulations

CASE 1 - 1.5 s.d. change

.	.05	.10	.15	.20	.25	.30
.2	10.48	9.08	7.97	8.20	8.06	7.66
.25	9.28	8.02	7.80	7.26	7.18	6.90
.3	8.32	7.20	7.05	6.58	6.56	6.37
.35	7.56	6.44	6.36	5.87	6.02	5.86
.4	7.25	6.26	5.94	5.80	5.53	5.44
.45	6.68	5.96	5.06	5.30	4.98	5.21
.5	6.36	5.54	5.14	4.64	4.74	4.74

CASE 2 - 2.0 s.d. change

.	.05	.10	.15	.20	.25	.30
.2	4.57	4.24	4.30	4.28	4.00	4.06
.25	4.49	4.09	3.94	3.96	3.81	3.93
.3	4.16	3.80	3.72	3.58	3.26	3.45
.35	3.69	3.64	3.48	3.20	3.28	3.24
.4	3.67	3.42	3.23	3.18	3.10	3.14
.45	3.51	3.25	3.17	2.98	2.94	3.06
.5	3.40	3.14	2.85	2.86	2.82	2.89

Cont.

CASE 3 - 3.0 s.d. change

.	.05	.10	.15	.20	.25	.30
.2	2.49	2.38	2.29	2.17	2.22	2.28
.25	2.34	2.27	2.24	2.18	2.23	2.19
.3	2.29	2.17	2.18	2.12	2.03	2.09
.35	2.18	2.16	2.12	2.06	2.09	2.00
.4	2.18	2.10	2.05	2.02	1.94	1.98
.45	2.14	2.10	1.95	1.93	1.93	1.94
.5	2.10	2.03	1.93	1.89	1.88	1.88

CASE 4 - 5 s.d. change

.	.05	.10	.15	.20	.25	.30
.2	1.68	1.59	1.58	1.57	1.57	1.56
.25	1.59	1.59	1.55	1.56	1.52	1.53
.3	1.55	1.56	1.54	1.55	1.53	1.52
.35	1.51	1.53	1.50	1.49	1.45	1.46
.4	1.50	1.44	1.48	1.50	1.49	1.43
.45	1.47	1.45	1.45	1.45	1.43	1.41
.5	1.47	1.47	1.44	1.40	1.39	1.37

TAB 3 - ERL / MA(1) MODEL (1000 simulations)

CASE 1 - TETA = .2

m = 0

.	.1	.2	.3
.2	63.78	55.90	
.3	40.36	35.34	
.4	29.32	23.85	

m = 1

	.1	.2	.3
	10.97	9.66	8.71
	7.85	7.77	7.38
	6.64	6.13	5.88

m = 1.5

.	.1	.2	.3
.2	4.62	4.40	4.03
.3	3.69	3.48	3.42
.4	3.12	3.00	2.82

m = 2

	.1	.2	.3
	2.41	2.39	2.34
	2.19	2.01	2.01
	1.98	1.90	1.90

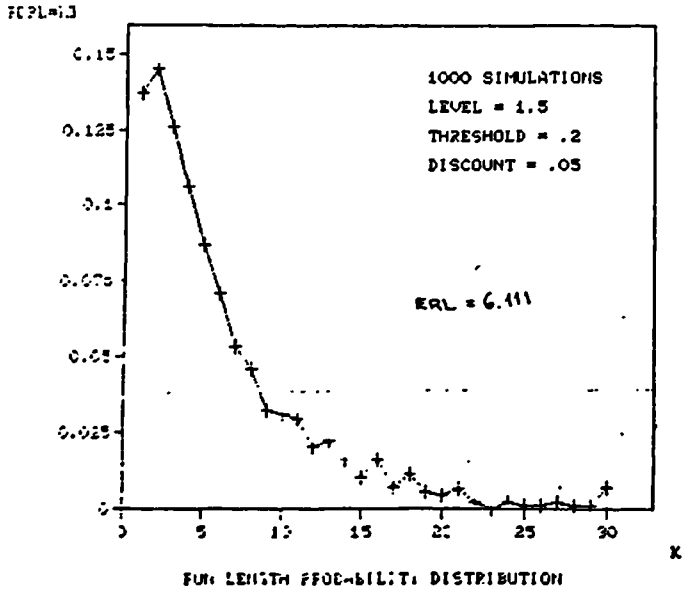
m = 3

.	.1	.2	.3
.2	1.34	1.34	1.35
.3	1.26	1.18	1.24
.4	1.17	1.16	1.16

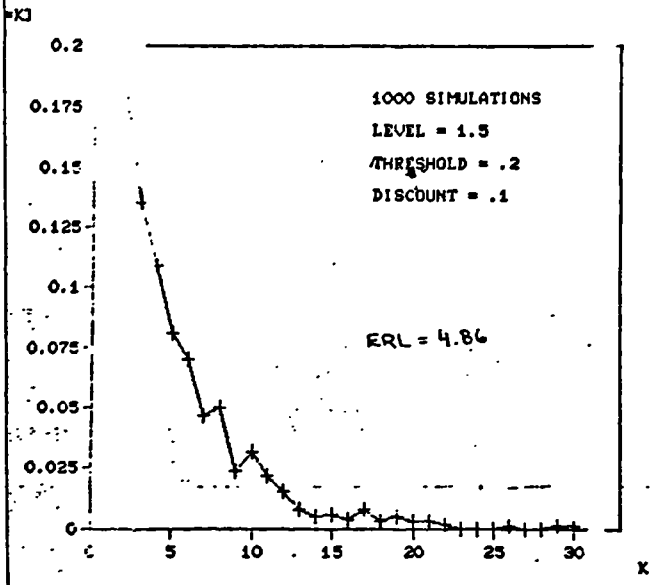
m = 5

	.1	.2	.3
	1.006	1.006	1.006
	1.004	1.002	1.004
	1.002	1.000	1.002

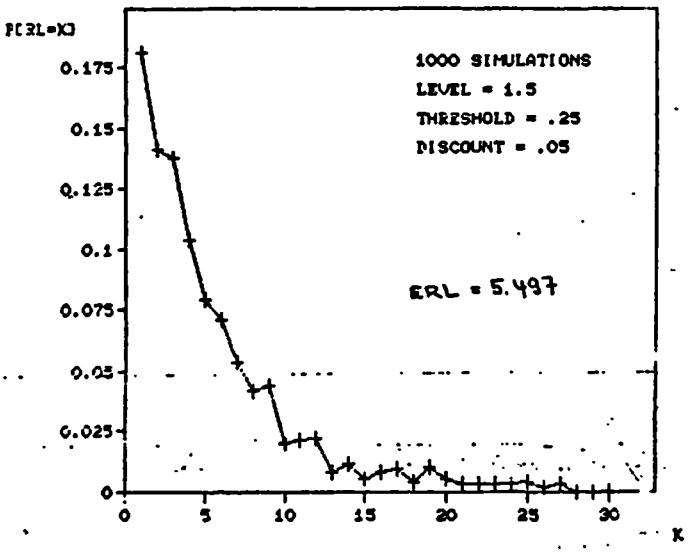
; RUN LENGTH PROBABILITY DISTRIBUTION



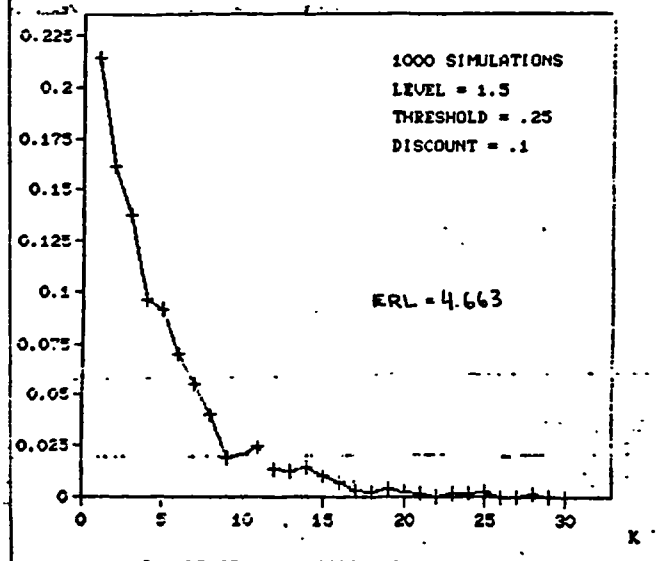
RUN LENGTH PROBABILITY DISTRIBUTION



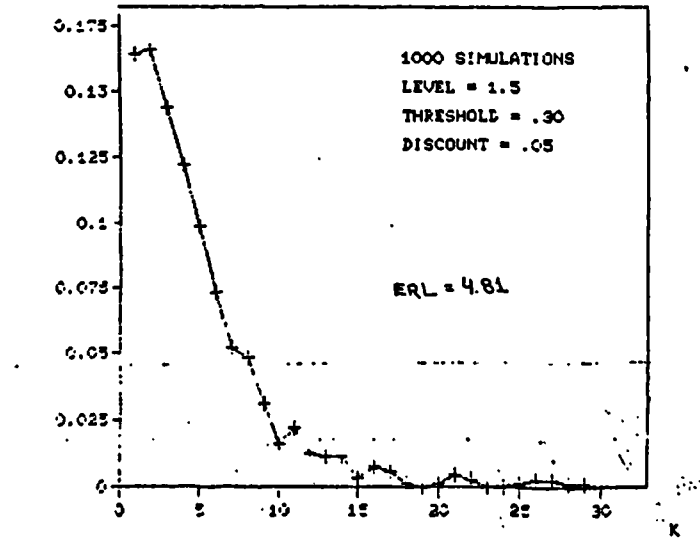
RUN LENGTH PROBABILITY DISTRIBUTION



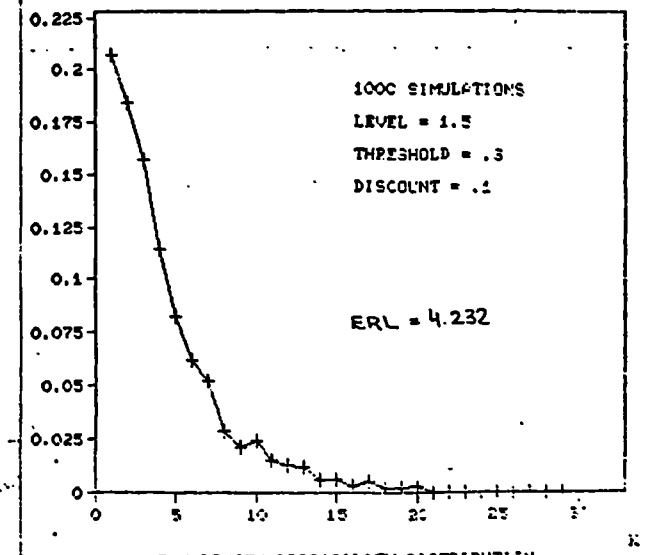
RUN LENGTH PROBABILITY DISTRIBUTION



RUN LENGTH PROBABILITY DISTRIBUTION



RUN LENGTH PROBABILITY DISTRIBUTION



RUN LENGTH PROBABILITY DISTRIBUTION

CHAPTER 5

MULTIVARIATE DYNAMIC LINEAR MODELS

5.1 - Introduction

The development of statistical procedures for modelling and analysis of vector time series is a very important theoretical issue with an enormous range of applications in many different areas. In practical situations such as in business, engineering, social or natural sciences, we frequently observe several related time series, and procedures for joint modelling, analysis or forecasting of such processes are necessary.

We commonly observe a fisical or economic process classified or disaggregated by geographical region or another factor, generating naturally a multivariate series of data. In such cases the correlation structure among the component series or, for instance, the joint probability distribution of one subset of series given values of another subset of series can be of extreme importance in a decision making or planning process.

The more general assumption we can make about these time processes is that they are non-stationary, and in a natural and Bayesian way we will consider here the case of dynamic Bayesian models - Harrison & West(1987) , West & Harrison(1989) or, more specifically, the class of Multivariate Dynamic Linear Models. In principle, the class of univariate (normal) dynamic linear models - D.L.M.'s for short - can be extended in a simple way, taking the observations at each time as vectors rather than scalars, providing a wide and rich class of models for multiple time series.

In fact, these models have already been defined since the original introduction of D.L.M.'s to statisticians by Harrison & Stevens (1976), and the basic theory is the same as in the univariate case, provided that we know the observational noise variance-covariance matrix V. This means that conditionally on the knowledge of V, the Bayesian estimation procedure for the process parameter THETA gives the same well known Kalman Filter-type equations used frequently in the univariate case , shown in chapter 2 .

This sort of unified DLM structure with respect to the dimension of the observations , in fact represents another extra advantage of the Markovian or state-space representation of time-series over more traditional forms of representation . While traditional time series analysis is primarily directed toward scalar-valued data , and usually represents time processes or their differenced version by scalar ARMA models , our DLM approach based on Markovian

representation of time series , treat several variables simultaneously as vector-valued variables .

In practice however, the observational noise variance-covariance matrix V is not known and some procedure for estimating it sequentially is necessary. This and related issues will be discussed throughout this chapter and some techniques for multivariate DLM analysis will be presented as well an analysis of its limitations. Also, we don't know the system noise covariance matrix W, but this is not a central issue here because the commonly used univariate technique of discount factors [Harrison & West(1987)] can be extended to the multivariate case in a straight forward way .

Alternatively, if we know the value of the system parameter THETA, a neat conjugate prior analysis to estimate the covariance matrix V is available, adopting an inverted-Wishart distribution for V . This, combined with the standard updating equations for estimating THETA would be perhaps the simplest ideas we could think of. However, as explained in section 5.2, although useful in some cases , this sort of conditional analysis does not constitute a general procedure and a more sensible one would be the joint estimation of both parameters . A natural candidate model is the multivariate normal-inverse Wishart distribution . Such basic initial ideas as well as the introduction of the notation will be considered in detail in the next section.

It is important to notice however, that for some special cases a full Bayesian solution is available. In the particular case of dimension one (univariate models) there is a neat conjugate prior analysis for estimating this observational noise variance (scalar) based on the inverted-gamma/normal distribution - Harrison & West(1987) - as shown in chapter 2.

Another important particular situation is the case of common components multivariate D.L.M.'s where each univariate marginal D.L.M. has the same F and G elements (regression and system matrix respectively) and all covariance matrices are scaled by the observational noise variance V in a Kronecker product fashion. In this case, a standard analysis based on the inverse-Wishart/matrix-Normal distribution - Dawid(1981) , Press(1982) - is presented by Quintana(1985).

These two particular cases are covered in section 5.3 of this chapter , where the common components limitations are analysed , and are special cases of the more general new methods we propose in the next chapter . Also , a new updating algorithm for the common components

multivariate DLM , based on Reference Analysis and Jeffreys'priors is presented in section 5.4 of this chapter , where an application with real data is provided in order to exemplify some of the common components drawbacks studied in section 5.3 . Finally , some technical results such as the Matrix Inversion Lemma , the matrix-normal / inverse-Wishart algorithm and some related results are presented in the form of Appendices to this chapter .

5.2 - PROBLEM FORMULATION AND GENERAL BACKGROUND

In this section we present formally the general multivariate D.L.M. model with the more simple possible forms of analysis in the first couple of sub-sections. It is followed in the next sub-section by a discussion of some aspects involved in a more general joint analysis .

5.2.1 The General Multivariate D.L.M.

This model was presented originally by Harrison & Stevens(1976) and is stated in terms of discrete intervals of time indexed by t .

DEFINITION : The multivariate (normal) D.L.M. for a vector of observations $\underline{y}_t$ of dimension d made at indexed times t=1,2,... is defined by the following equations:

$$\text{i) observation equation :} \qquad \underline{y}_t = F_t^T.\underline{\theta}_t + \underline{v}_t, \quad \underline{v}_t \sim N(\underline{0}, V_t) \tag{5.1}$$

$$\text{ii) system equation:} \qquad \underline{\theta}_t = G_t.\underline{\theta}_{t-1} + \underline{w}_t, \quad \underline{w}_t \sim N(\underline{0}, W_t) \tag{5.2}$$

where:

$\underline{\theta}_t$ is a px1 vector of process parameters at time t.

F_t is a pxd matrix of (known) constants and/or independent regressors.

G_t is a pxp known system matrix .

$\underline{v}_t$ & $\underline{w}_t$ are sequences of independent zero mean normal random vectors which without loss in generality are also independent of each other .

V_t is a dxd (unknown) observational noise variance-covariance matrix.

W_t is a pxp (known) system noise variance-covariance matrix.

Comments : a) The variance-covariance matrix structure :

i) the unknown observational variance V_t can vary in time , but usually varies much slower than the process parameter $\underline{\theta}_t$, or it is constant up to perhaps a scale factor . Another

possibility of variance dynamics scheme is the use of power or other laws , which , for the univariate case were discussed briefly in chapter 2 .

ii) In practice the unknown system noise variance-covariance matrix W_t will be specified through the use of a vector of discount factors b in a simple extension of the technique used for univariate D.L.M.'s as in Harrison & West (1987).

b) Some special particular cases :

i) Static multivariate regression models ($W_t = 0$; $G_t = I$) . In this class of models we have , among others , for instance , the general linear model with common regression coeficients - Box & Tiao (1973) and the seemingly unrelated regression equations model - Zellner(1971) , as well general simultaneous equations models .

ii) Among s ochastic process models , we could mention as particular cases for instance , the multidimensional random walk ($V_t = 0$; $F_t = G_t = I$) , or even more general Markovian processes with $G_t \neq I$. Among time series models , we have as particular case the Common Components Models ($\mathbf{F}_t = I \otimes F_t$; $\mathbf{G}_t = I \otimes G_t$; $\mathbf{W}_t = V_t \otimes W_t$) , as for instance the dynamic linear matrix-variate model - Quintana(1985) , Quintana & West(1987) , and the generalized multivariate exponential smoothing model - Harvey(1986) , among others .

c) NOTATION : Initially , lets consider the following notation ,

$D_t = (\underline{y}_t, D_{t-1})$ represents all information available (data and others) about the process at time t.

$\underline{\theta}_t / V_t, D_{t-1} \sim N(\underline{a}_t, R_t)$ is the prior distribution for the process parameter at time t , conditional on V_t .

$\underline{\theta}_t / V_t, D_t \sim N(\underline{m}_t, C_t)$ is the posterior distribution for $\underline{\theta}_t$, conditional on V_t .

5.2.2 - Basic Conditional Analysis :

a) Conditioned on the value of V_t , we can learn about $\underline{\theta}_t$ in a standard Bayesian fashion (Conjugate Prior Analysis) as follows:

i) time updating : $(\underline{\theta}_t / D_{t-1}) \sim N(\underline{a}_t, R_t)$ where:

$$\underline{a}_t = G_t . \underline{m}_{t-1} \tag{5.3a}$$

$$R_t = G_t . C_{t-1} . G_t^T + W_t \tag{5.3b}$$

where W_t is specified through a given discount vector $\underline{b}$.

ii) observation updating: $(\underline{\theta}_t/D_t) \sim N(\underline{m}_t, C_t)$ where:

$$\underline{m}_t = \underline{a}_t + A_t.(\underline{y}_t - F_t^T.\underline{a}_t) \tag{5.4a}$$

$$C_t = R_t - A_t.Q_t.A_t^T \tag{5.4b}$$

$$A_t = R_t.F_t.Q_t^{-1} \tag{5.4c}$$

$$Q_t = F_t^T.R_t.F_t + V_t \tag{5.4d}$$

iii) initial information: $(\underline{\theta}_0/D_0) \sim N(\underline{m}_0, C_0)$

a1) Reference Analysis : Alternatively , when there is practically no initial information available about $\underline{\theta}_t$ such as $(\underline{m}_0, C_0)$ or when it is difficult to setup such initial values , it is still possible to carry on a special Bayesian analysis (Reference Analysis) for the multivariate D.L.M. When V_t is known and we consider a vague or non-informative initial prior distribution for $\underline{\theta}$, the updating equations for the multivariate D.L.M. can be obtained from the previous algorithm of this section using the Binomial Inverse Theorem or Matrix Inverse Lemma (see Appendix 5.1 of this chapter) as follows :

i) From the equations 5.4b and 5.4c , we have ,

$$C_t = R_t - A_t.Q_t.A_t^T = R_t - R_t.F_t.(F_t^T.R_t.F_t + V_t)^{-1}.F_t^T.R_t$$

or equivalently , considering the Matrix Inversion Lemma ,

$$C_t^{-1} = R_t^{-1} + F_t.V_t^{-1}.F_t^T \qquad \text{or} \qquad K_t = H_t + F_t.V_t^{-1}.F_t^T \tag{5.5a}$$

where $H_t = R_t^{-1}$ and $K_t = C_t^{-1}$ are respectively the prior and posterior Inverse Covariance or Information matrix associated with the process parameter $\underline{\theta}_t$ when such inverses exist · otherwise we update K_t and H_t without such interpretation . Also , from 5.3b , $R_t^{-1} = (W_t + G_t.C_{t-1}.G_t^T)^{-1}$, or

$$H_t = W_t^{-1} - W_t^{-1}.G_t.P_t^{-1}.G_t^T.W_t^{-1} \tag{5.5b}$$

$$P_t = G_t^T.W_t^{-1}.G_t + K_{t-1} \tag{5.5c}$$

and the cycle for updating the information matrices is complete .

ii) Now , from equations 5.4c and 5.4d , $A_t = R_t.F_t^T.(F_t^T.R_t.F_t + V_t)^{-1}$ and , by the Matrix Inversion Lemma , we have

$$A_t = (F_t^T.V_t^{-1}.F_t + R_t^{-1})^{-1}.F_t^T.V_t^{-1} \qquad \text{or} \qquad A_t = C_t.F_t^T.V_t^{-1} \tag{5.6a}$$

Now , from (5.4a) we have , $\underline{m}_t = (I - A_t.F_t).\underline{a}_t + A_t.\underline{y}_t$, or

$$C_t^{-1}.\underline{m}_t = (C_t^{-1} - F_t^T.V_t^{-1}.F_t).\underline{a}_t + F_t^T.V_t^{-1}.\underline{y}_t \qquad \text{which gives}$$

$$\underline{k}_t = \underline{h}_t + F_t^T.V_t^{-1}.\underline{y}_t \tag{5.6b}$$

where $\underline{h}_t = R_t^{-1}.\underline{a}_t$ and $\underline{k}_t = C_t^{-1}.\underline{m}_t$. Also , from (5.3a) we have , $\underline{h}_t = R_t^{-1}.G_t.\underline{m}_{t-1} = R_t^{-1}.G_t.C_{t-1}.\underline{k}_{t-1}$ or, substituting R_t^{-1} and C_{t-1} for equivalent expressions , we get

$$\underline{h}_t = (W_t^{-1} - W_t^{-1}.G_t.P_t^{-1}.G_t^T.W_t^{-1}).G_t.(P_t^{-1} + P_t^{-1}.G_t.U_t^{-1}.G_t^T.P_t^{-1}).\underline{k}_{t-1}$$

where $U_t^{-1} = (W_t.R_t^{-1}.W_t)^{-1} = (W_t - G_t.P_t^{-1}.G_t^T)^{-1}$. Then , after some algebra , we get finally

$$\underline{h}_t = W_t^{-1}.G_t.P_t^{-1}.\underline{k}_{t-1} \tag{5.6c}$$

The equations (5.5a) to (5.6c) with the Initial Values $H_1 = 0, \underline{h}_1 = 0$ define the so called Information or Inverse-Covariance Filter . These updating equations coincide with the Reference Analysis of the multivariate D.L.M. , presented in the Appendix 5.1 .

b)Conditional on the value of $\underline{\theta}_t$, we can learn about V_t using standard conjugate prior analysis for the model (5.1)-(5.2) as follows:

i) prior distribution : $(V_t/D_{t-1}) \sim W^{-1}[S_{t-1}; n_{t-1}]$, where S_{t-1} and n_{t-1} are respectively the shape parameter and d.f. of the inverse-Wishart distribution.

ii) posterior distribution: $(V_t/D_t) \sim W^{-1}[S_t; n_t]$ where:

$$S_t = S_{t-1} + \underline{\nu}_t.\underline{\nu}_t^T \tag{5.7a}$$

$$\underline{\nu}_t = \underline{y}_t - F_t^T.\underline{\theta}_t \tag{5.7b}$$

$$n_t = n_{t-1} + 1 \tag{5.7c}$$

$$\hat{V}_t = E(V_t/D_t) = n_t^{-1}.S_t \tag{5.7d}$$

5.2.3 - Toward a Joint Analysis :

A possible way to estimate both $\underline{\theta}_t$ and V_t would be, at each time t, to iterate between these two conditional estimators - given by the equations (5.3a)-(5.4d) and (5.7a)-(5.7d) - until we get some kind of convergence, and then go to the next time t+1. However, this procedure can be very time demanding if we don't get convergence in one or two iterations, and more efficient methods are necessary. Otherwise, if we pre-specify in one iteration every time, the performance can be very poor because we are not taking fully into account the uncertainty about one parameter when we estimate the other one and vice-versa.

On the other hand, a joint estimator for $\underline{\theta}_t$ and V_t can be built from the joint posterior distribution of these parameters. More specifically, a multivariate normal/inverse-Wishart prior density for the model given by the equations (5.1)-(5.2) will give us :

$$\begin{aligned} p(\underline{\theta}_t, V_t/D_t) &\propto p(\underline{\theta}_t, V_t/D_{t-1}).p(\underline{y}_t/\underline{\theta}_t, V_t) \\ &\propto p(V_t \ \ D_{t-1}).p(\underline{\theta}_t/V_t, D_{t-1}).p(\underline{y}_t/\underline{\theta}_t, V_t) \\ p(\underline{\theta}_t, V_t/D_t) &\propto |V_t|^{-\frac{n_t-d-1}{2}}.\exp[-\frac{1}{2}.trS_{t-1}.V_t^{-1}] \\ \exp[-\frac{1}{2}(\underline{\theta}_t - \underline{a}_t)^T.R_t^{-1}.(\underline{\theta}_t - \underline{a}_t)].|V_t|^{-\frac{1}{2}}.\exp[-\frac{1}{2}(\underline{y}_t - F_t.\underline{\theta}_t)^T.V_t^{-1}.(\underline{y}_t - F_t.\underline{\theta}_t)] \end{aligned} \tag{5.8}$$

where the three factors represent respectivelly the prior marginal distribution for V_t , the (conditional on V_t) prior distribution for $\underline{\theta}_t$ and the likelihood for $\underline{\theta}_t$ and V_t . Of course, it would not be easy or tractable to integrate out (5.8) in order to get joint posterior moments for $\underline{\theta}_t$ and V_t .However, joint modes $(\underline{\theta}_t^*, V_t^*)$, which involve only derivatives can be more immediately obtained, resulting in the following modal equations:

$$\underline{\theta}_t^* = \underline{a}_t + R_t.F_t^T.[F_t^T.R_t.F_t + V_t^*]^{-1}.(\underline{y}_t - F_t.\underline{a}_t) \tag{5.9a}$$

$$V_t^* = (n_{t-1} - d)^{-1}.[V_{t-1}^* + (\underline{y}_t - F_t.\underline{\theta}_t^*).(\underline{y}_t - F_t.\underline{\theta}_t^*)^T] \tag{5.9b}$$

These equations should be solved iteratively to provide the values of the modes for use as point estimates of the parameters of our sequential normal-Wishart prior specification 1 step on . But the difficulty is the same as before : an iterative solution is not computationally attractive and simple approximations supposing independence between the two equations are not good estimators.

This sort of problem arises in general because there is not a set of tractable sufficient statistics , and it is interesting to mention that this sort of difficulty in estimating noise variances is not restricted to dynamic models . It happens for instance in static linear models such as the general linear model with common regression coeficients - Box & Tiao(1973) and the seemingly unrelated regression model - Zellner(1971) . Also, in dynamic models under constant noise variances following an Inverse-Wishart distribution , the posterior distribution of $\underline{\theta}_t$ is an intractable multivariate poly-t distribution (Broemeling , 1985) . It is clear now that extra assumptions are necessary in order to obtain a tractable procedure for on-line variance learning .

After we have this initial exploration into the nature of the problem we continue our investigation as follows . In the next section we present two particular cases where a fully bayesian solution is possible. After a brief review of the univariate case that provide us with some important insight , we show the serious limitations of the common component multivariate D.L.M.

5.3 - CONJUGATE ANALYSIS FOR SPECIAL D.L.M.'s

5.3.1 - The Univariate Case : In the case where V_t is an unknown scalar but constant ($V_t = V$) there is a fully conjugate bayesian learning procedure. Here we present briefly this analysis for future reference . Further details and references can be found in Harrison & West (1987) or in chapter 2 of this thesis , where we present also other alternative algorithms for DLM analysis .

In our case of $V_t = V$ constant for all t and d=1 , the model (5.1)-(5.2) can be written as:

$$y_t = F_t^T . \underline{\theta}_t + v_t, \qquad v_t \sim N[0, V] \tag{5.10}$$

$$\underline{\theta}_t = G_t . \underline{\theta}_{t-1} + \underline{w}_t, \qquad \underline{w}_t \sim N[\underline{0}, V.W_t^*] \tag{5.10a}$$

The analysis is as follows:

i) At time t-1 the variance V is modelled by an inverse chi-square distribution with n_{t-1} d.f. and point estimate S_{t-1} , or

$$\left(\frac{n_{t-1} . S_{t-1}}{V} / D_{t-1} \right) \sim \chi^2_{n_{t-1}} \tag{5.11}$$

and the parameter vector $\underline{\theta}_t$ has a prior distribution conditional on V given by,

$$(\underline{\theta}_t/V, D_{t-1}) \sim N(\underline{a}_t, R_t) \tag{5.11a}$$

where $\underline{a}_t$ and G_t follow the same time evolution given respectively by the equations 5.3a and 5.3b , and the corresponding marginal for $\underline{\theta}_t$ is a multivariate t distribution with n_{t-1} d.f.

ii) The posterior distribution for the unknown parameters will be also a normal / inverse chi-square distribution with parameters n_t & S_t for V and $\underline{m}_t$ & C_t for $\underline{\theta}_t$ given by:

$$S_t = S_{t-1}.(n_{t-1} + e_t^2/Q_t)/n_t \tag{5.12a}$$

$$n_t = n_{t-1} + 1 \tag{5.12b}$$

$$e_t = y_t - F^T.\underline{a}_t \tag{5.12c}$$

$$Q_t = F_t^T.R_t.F_t + S_{t-1} \tag{5.12d}$$

and also

$$\underline{m}_t = \underline{a}_t + A_t.e_t \tag{5.13a}$$

$$C_t = (R_t - A_t.Q_t.A_t^T)S_t/S_{t-1} \tag{5.13b}$$

$$A_t = R_t.F_t.Q_t^{-1} \tag{5.13c}$$

Comments:

i) It is important to notice that the key difference between the case of V known with posterior distribution given by the equations (5.4) and the present case of unknown variance with posterior distribution given by the equations (5.13) is the presence of the scale factor S_t/S_{t-1} (equation 5.13b) correcting the standard expression for the posterior variance C_t given by the equation 5.4b . This idea of variance scaling will be very useful in order to appreciate better some aspects of the techniques proposed in the next chapter .

ii) An alternative to the conjugate prior analysis of the univariate D.L.M. , given by equations 5.12a - 5.13c , is to consider a vague or non-informative initial prior . This sort of Reference Analysis version of the previous algorithm as well some implementation aspects are covered in detail in chapter 2 of this thesis .

iii) This case of a constant unknown variance V is readily extended to that of general variance laws , which are defined up to an unknown constant (West & Harrison , 1989) .

5.3.2 The Common Components Model

This model is presented in Quintana(1985,87), developed in Quintana& West (1986,88) and given in West & Harrison (1989) . Also , a non-Bayesian version of this model is presented in Harvey(1986) . The CCM is a particular case of the multivariate D.L.M. model of section 2.1 and its essential feature is that all marginal univariate component DLM's have the same F,G & W elements.

In order to investigate its properties and compare with other methods, we present formally the model here. This is done, rewriting the multivariate DLM equations (5.1)-(5.2), adding the following assumptions:

i) the design elements (regression and system matrices) $\mathbf{F}_t$ & $\mathbf{G}_t$ are specified respectively by

$$\mathbf{F}_t = (I_d \otimes F_t) \qquad \& \qquad \mathbf{G}_t = (I_d \otimes G_t)$$

where the regression and system matrices, respectively F_t & G_t are common for all the d univariate marginal D.L.M. components.

ii) all the variance-covariance matrix structure is scaled by the observational variance V in a Kronecker product fashion ,

$$\mathbf{W}_t = V \otimes W_t \qquad \& \qquad \mathbf{C}_t = V \otimes C_t$$

where the corresponding univariate elements W_t and C_t are common for all the d univariate marginal D.L.M. components . For simplicity , we consider V as constant although it can readily be considered such that its elements each follow variance laws known up to proportionality constants .

iii) the observational variance V given D_{t-1} follows an inverse-Wishart matrix distribution,

$$V/D_{t-1} \sim W^{-1}[\,d_{t-1}, n_{t-1}\,]$$

where d_{t-1} is the shape parameter and n_{t-1} is the d.f. (d_{t-1}/n_{t-1} is a point estimate of V/D_{t-1}).

This gives the common components multivariate D.L.M. defined by the following equations:

$$\text{i) observation equation:} \quad \underline{y}_t = (I_d \otimes F_t^T).\underline{\theta}_t + \underline{v}_t, \qquad \underline{v}_t \sim N(\underline{0}, V) \tag{5.14}$$

$$\text{ii) system equation:} \quad \underline{\theta}_t = (I_d \otimes G_t).\underline{\theta}_{t-1} + \underline{w}_t, \quad \underline{w}_t \sim N(\underline{0}, V \otimes W_t) \tag{5.15}$$

iii) prior information: $(\underline{\theta}_{t-1}/V, D_{t-1}) \sim N(\underline{m}_{t-1}, V \otimes C_{t-1})$ (5.16)

$$(V/D_{t-1}) \sim W^{-1}(d_{t-1}, n_{t-1}) \tag{5.16a}$$

This model can be equivalently formulated in terms of matrix normal notation , which enables an efficient updating algorithm , that is shown in the Appendix 5.2

However, in order to understand better the characteristics and limitations of this model we present here the following Model Analysis , where , for simplicity , we consider F_t and G_t as constant .

i) By construction, all covariance matrices are scaled by V, and the joint prior distribution of observations and state parameters will be given by

$$\begin{pmatrix} \underline{\theta}_t \\ \underline{y}_t \end{pmatrix} \Big| \; D_{t-1}, V \sim N\left\{ \begin{pmatrix} \underline{a}_t \\ \underline{f}_t \end{pmatrix}; \begin{pmatrix} V \otimes R_t & V \otimes R_t.F \\ . & V \otimes Q_t \end{pmatrix} \right\} \quad \text{where} \tag{5.17}$$

$$\underline{a}_t = (I \otimes G).\underline{m}_{t-1} \quad ; \quad \underline{f}_t = (I \otimes F^T).\underline{a}_t \tag{5.17a}$$

$$Q_t = F^T.R_t.F + 1 \quad ; \quad R_t = G.C_{t-1}.G^T + W_t \tag{5.17b}$$

Also, as a consequence ,

$$Cov(\underline{\theta}_t, \underline{y}_t).[Var(\underline{y}_t)]^{-1} = I \otimes A_t \tag{5.17c}$$

where $A_t = R_t.F.Q_t^{-1}$

ii) Consequently, the posterior distribution for the state parameter and observational variance will be given by:

$$(\underline{\theta}_t/D_t, V) \sim N(\underline{m}_t; V \otimes C_t) \tag{5.18}$$

where :

$$\underline{m}_t = \underline{a}_t + (I \otimes A_t).(\underline{y}_t - \underline{f}_t) \quad \text{and} \quad C_t = R_t - A_t.Q_t.A_t^T \tag{5.18a}$$

Also, $(V/D_t) \sim W^{-1}(d_t; n_t)$ where,

$$d_t = d_{t-1} + \underline{e}_t.\underline{e}_t^T.Q_t^{-1} \quad ; \quad n_t = n_{t-1} + 1 \quad ; \quad \underline{e}_t = \underline{y}_t - \underline{f}_t \tag{5.18b}$$

Important : From the above analysis , it is clear that the model is very restrictive and limited in application . The far too strong structural model assumptions result in :

i) The sequence of updatings of the mean $E(\underline{\theta}_t/D_t)$ is totally independent of the variance-covariance matrix V . This is clear since A_t does not depend on V. Or, in other words, we

don't have a general fully joint estimation procedure because the mean updating doesn't take into account the observational variance . Consequently , the location estimates do not benefit from any estimate of the variance and , as we have discussed before in section 5.1 , this is a restrictive and very bad property.

ii) The marginal forecast distributions of each series and the marginal distributions of their parameter blocks are identical to those derived simply by operating with the series independently with individual models. Again, it means that we don't have a general multivariate model, but only a combination of a set of univariate individual normal DLM's.

Writing $\underline{y}_t' = (y_{1t}, ..., y_{dt})$ and $\underline{\theta}_t' = (\underline{\theta}_{1t}', ..., \underline{\theta}_{pt}')$, the forecast distributions $y_{i,t+k}/D_0, y_i^t$ and $\underline{\theta}_{i\ t+k}/D_0, y_i^t$ are identical to those derived from the commensurate individual DLM applied just to the i^{th} observation series . Thus all forecast and retrospective joint distributions conditional on all $\underline{y}_t$ to time T are identical to those conditional on $y_{i,t}$ to time T alone (T = 1,2,3,..) . This arises since if $\underline{e}_t' = (e_{1t}, .., e_{dt})$ and $d_t = \{d_{i,t}\}$, we have

$$d_{i,t} = d_{i,t-1} + e_{it}^2 . Q_t^{-1}$$

and of course Q_t is independent of V .

iii) It requires common marginal models with , in general , not only the pair F_t, G_t the same for each component series but a common 'generalized signal-to- noise ratio'. In fact , if we call $\underline{\mu}_t = \mathbf{F}_t^T . \underline{\theta}_t$ of signal and $\underline{v}_t$ of noise , the prior generalized signal-to-noise ratio (of variances) for the common components multivariate DLM will be given by ,

$$\begin{aligned} Var[\underline{\mu}_t / D_{t-1}] &= Var[\mathbf{F}_t^T . \underline{\theta}_t / D_{t-1}] = \mathbf{F}_t^T . \mathbf{R}_t . \mathbf{F}_t \\ &= (I \otimes F_t^T).(V \otimes R_t).(I \otimes F_t) = V \otimes F_t^T . R_t . F_t \end{aligned}$$

Then , the generalized ratio of $V(\underline{\mu}_t)$ by $V(\underline{v}_t)$ is a scalar factor common to all univariate marginal models given by $\lambda_t = F_t^T . R_t . F_t$. In a similar way , the posterior generalized signal-to- noise ratio of variances is the scalar $F_t^T . C_t . F_t$ common to all univariate marginal component models , which is a very restrictive property .

iv) Effective intervention is virtually impossible within the model since it is not possible to retain the $V \otimes C_t$ structure and alter the relative parameter uncertainties with a common C_t

.

v) Extensions to include non-linearities in the observation or system equation and usual linearisations via Taylor series are not compatible with the common component structure , as recognised by Quintana(1987) , pg 116 .

vi) Since the system matrix has the form $\mathbf{G}_t = (I \otimes G)$ with the block G common to all marginal series , it is not possible the modelling of shared components . That is the case , for instance , of a model with the same seasonal parameters for all series . If we write such model (artificially and non-parsimoniously) as a CCM , by repeating the shared component to all series , the only way to guarantee that all these repeated parameters are the same is to impose a linear parametric constraint to the model ($L.\underline{\theta}_t = \underline{0}$ for all t , where L is a contrast matrix of zeros and ones) . It happens that the CCM variance structure is not compatible with such linear constraint , and therefore , the modelling of any shared component is not compatible with the CCM framework .

Thus , the common components model is practically worthless . The restriction that F and G must be the same and that the signal-to-noise ratio is also the same across all components is far too particular . And when this prevents effective interventions and does not lead to any differences in the marginal distributions of the individual series from those derived using univariate methods , it is clearly seen that the value of this class of models is extremely limited .

In order to overcome these serious limitations we propose a more general methodology in the next chapter where such restrictions do not apply.

However , in order to make the Bayesian analysis of the common components multivariate DLM as complete as possible , and also to avoid the eventual inconvenience associated with prior specification , we provide an alternative new updating algorithm for this model in the next section where reference priors (Jeffreys'priors) are used . This algorithm will be used in an application with real data , in order to stress and exemplify in detail some of the drawbacks just mentioned .

5.4 - Reference Analysis of Common Components Multivariate DLM's

5.4.1 - Reference Analysis theory

Theorem 5.1 : For the Common Components Multivariate DLM defined by the equations

(5.38)-(5.39) , let the initial prior information be represented by the reference form

$$p(\Theta_t, V/D_t) \propto |V|^{-\frac{1}{2}(d+1)}$$

Then , assuming that W_t in (5.39) has full rank , we have :

A) The joint prior and posterior distributions of the matrix process parameter Θ_t and observation noise variance-covariance matrix V_t at time $t = 1, 2, ..$ are given by

$$p(\Theta_t, V/D_{t-1}) \propto |V|^{-\frac{1}{2}.(a_{t-1}+d+1)}.exp\{-\frac{1}{2}.trV^{-1}.[\Theta_t^T.H_t.\Theta_t - 2.\Theta_t^T.T_t + L_t]\}$$

$$p(\Theta_t, V/D_t) \propto |V|^{-\frac{1}{2}(a_t+d+1)}.exp\{-\frac{1}{2}.trV^{-1}.[\Theta_t^T.K_t.\Theta_t - 2.\Theta_t^T.U_t + E_t]\}$$

where :

$$H_t = W_t^{-1} - W_t^{-1}.G_t.Z_t^{-1}.G_t.W_t^{-1} \tag{5.19}$$

$$Z_t = G_t^T.W_t^{-1}.G_t + K_{t-1} \tag{5.19a}$$

$$T_t = W_t^{-1}.G_t.Z_t^{-1}.U_{t-1} \tag{5.20}$$

$$K_t = H_t + F_t.F_t^T \tag{5.21}$$

$$U_t = T_t + F_t.\underline{y}_t \tag{5.22}$$

$$L_t = E_{t-1} - U_{t-1}^T.Z_t^{-1}.U_{t-1} \tag{5.23}$$

$$E_t = L_t + \underline{y}_t.\underline{y}_t^T \tag{5.24}$$

$$a_t = a_{t-1} + 1 \tag{5.25}$$

with $H_1 = 0, T_1 = 0, L_1 = 0$ and $a_1 = 0$ as initial setup .

B) For $t \geq t_P$, where $t_P = p + \frac{1}{2}.d.(d+1)$ (if there is missing data in these first observations or problems of collinearity if the model includes regressors then , t_P should be increased properly) , the posterior distribution for $(\Theta_t, V/D_t)$ is a matrix-normal / inverse-Wishart distribution with parametrization (M_t, C_t, S_t, n_t) given by

$$M_t = K_t^{-1}.U_t \quad , \quad C_t = K_t^{-1} \quad , \quad S_t = E_t - U_t.M_t \quad \text{and} \quad n_t = a_t - d \tag{5.26}$$

Proof :

A) The proof is by finite induction on t as follows : Assume that the given joint prior distribution for (Θ_t, V) is true for t-1 ,

$$p(\Theta_t, V/D_{t-1}) \propto |V|^{-\frac{1}{2}(a_{t-1}+d+1)}.exp\{-\frac{1}{2}.trV^{-1}.(\Theta_t^T.H_t.\Theta_t - 2.\Theta_t^T.T_t + L_t)\}$$

From the model , the likelihood of an observation $\underline{y}_t$ is given by ,

$$l(\Theta_t, V/\underline{y}_t) \propto |V|^{-\frac{1}{2}}.exp\{-\frac{1}{2}.trV_t^{-1}.(\underline{y}_t - F_t^T.\Theta_t).(\underline{y}_t - F_t^T.\Theta_t)^T\}$$
$$\propto |V|^{-\frac{1}{2}}.exp\{-\frac{1}{2}.trV^{-1}.(\Theta_t^T.F_t.F_t^T.\Theta_t - 2.\Theta_t^T.F_t.\underline{y}_t + \underline{y}_t.\underline{y}_t^T)\}$$

By Bayes' Theorem , the joint posterior distribution is given by ,

$$p(\Theta_t, V/D_t) \propto p(\Theta_t, V/D_{t-1}).l(\Theta_t, V/\underline{y}_t)$$
$$\propto |V|^{-\frac{1}{2}(a_t+d+1)}.exp\{-\frac{1}{2}.trV^{-1}.(\Theta_t^T.K_t.\Theta_t - 2.\Theta_t^T.U_t + E_t)\}$$

where

$$K_t = H_t + F_t.F_t^T$$
$$U_t = T_t + F_t.\underline{y}_t$$
$$E_t = L_t + \underline{y}_t.\underline{y}_t^T$$
$$a_t = a_{t-1} + 1$$

Now , the joint prior distribution at time t+1 , will be given by ,

$$p(\Theta_{t+1}, V/D_t) = \int p(\Theta_{t+1}, V/\Theta_t, D_t).p(\Theta_t/D_t).d\Theta_t$$
$$= \int p(\Theta_{t+1}/\Theta_t, V, D_t).p(V/\Theta_t, D_t).p(\Theta_t/D_t).d\Theta_t$$
$$= \int p(\Theta_{t+1}/\Theta_t, V, D_t).p(\Theta_t, V/D_t).d\Theta_t$$

From the model system equation , the first term in the integral is the matrix- normal distribution N ($G_{t+1}.\Theta_t, W_{t+1}, V$) , and then

$$p(\Theta_{t+1}, V/D_t) \propto \int V|^{-\frac{1}{2}}.exp\{-\frac{1}{2}.trV^{-1}.(\Theta_{t+1} - G_{t+1}.\Theta_t)^T.W_{t+1}^{-1}.(\Theta_{t+1} - G_{t+1}.\Theta_t)\}.$$
$$.|V|^{-\frac{1}{2}.(a_t+d+1)}.exp\{-\frac{1}{2}.trV^{-1}.(\Theta_t^T.K_t.\Theta_t - 2.\Theta_t^T.U_t + E_t)\}.d\Theta_t$$
$$\propto \int |V|^{-\frac{1}{2}.(a_t+2d+1)}.exp\{-\frac{1}{2}.trV^{-1}.(\Theta_t^T.Z_t.\Theta_t - 2.\Theta_t^T.\Gamma_t + B_t)\}.d\Theta_t$$
$$\propto \int |V|^{-\frac{1}{2}.(a_t+2d+1)}.exp\{-\frac{1}{2}.trV^{-1}.(\Theta_t - \alpha_t)^T.Z_t.(\Theta_t - \alpha_t) + R_t\}.d\Theta_t$$

where

$$Z_t = K_t + G_{t+1}^T.W_{t+1}^{-1}.G_{t+1}$$
$$\alpha_t = Z_t^{-1}.\Gamma_t = Z_t^{-1}.(U_t + G_{t+1}^T.W_{t+1}^{-1}.\Theta_{t+1})$$
$$B_t = E_t + \Theta_{t+1}^T.W_{t+1}^{-1}.\Theta_{t+1}$$
$$R_t = B_t - \alpha_t^T.Z_t.\alpha_t$$

Integrating out the matrix-normal density , we get finally

$$p(\Theta_{t+1}, V/D_t) \propto |V|^{-\frac{1}{2}.(a_t+d+1)}.exp\{-\frac{1}{2}.R_t\}$$

Now , substituting B_t and α_t by their expressions in R_t , we get

$$\begin{aligned} R_t &= \Theta_{t+1}^T.W_{t+1}^{-1}.\Theta_{t+1} - (U_t + G_{t+1}^T.W_{t+1}^{-1}.\Theta_{t+1})^T.Z_t^{-1}.(U_t + G_{t+1}^T.W_{t+1}^{-1}.\Theta_{t+1}) + E_t \\ &= \Theta_{t+1}^T.H_{t+1}.\Theta_{t+1} - 2.\Theta_{t+1}^T.T_{t+1} + L_{t+1} \end{aligned}$$

where

$$\begin{aligned} H_{t+1} &= W_{t+1}^{-1} - W_{t+1}^{-1}.G_{t+1}.Z_t^{-1}.G_{t+1}^T.W_{t+1}^{-1} \\ T_{t+1} &= W_{t+1}^{-1}.G_{t+1}.Z_t^{-1}.U_t \\ L_{t+1} &= E_t - U_t^T.Z_t^{-1}.U_t \end{aligned}$$

Finally , for $t = 1$, the theorem is also valid , since if we set $H_1 = 0, T_1 = 0, L_1 = 0$ and $a_1 = 0$, we get from the expression for the joint prior distribution , exactly the reference form

$$p(\Theta_1, V/D_0) \propto |V^{-\frac{1}{2}.(d+1)}$$

B) From the part A of this theorem , the joint posterior is given by

$$p(\Theta_t, V/D_t) \propto V|^{-\frac{1}{2}.(a_t+d+1)}.exp\{-\frac{1}{2}.trV^{-1}.(\Theta_t^T.K_t.\Theta_t - 2.\Theta_t^T.U_t + E_t)\}$$

which for $t \geq t_P$ (when the posterior distributions are proper) can by rewritten as ,

$$\begin{aligned} p(\Theta_t, V/D_t) &\propto V|^{-\frac{1}{2}.(a_t+d+1)}.exp\{-\frac{1}{2}.trV^{-1}.[(\Theta_t - M_t)^T.K_t.(\Theta_t - M_t) + E_t - M_t^T.K_t.M_t]\} \\ &\propto |V|^{-\frac{1}{2}(a_t+d+1)} exp\{-\frac{1}{2}trV^{-1}(E_t - U_t^T.M_t)\} exp\{-\frac{1}{2}trV^{-1}(\Theta_t - M_t)^T.K_t(\Theta_t - M_t)\} \end{aligned}$$

where M_t and C_t are identified from the matrix-normal component as

$$M_t = K_t^{-1}.U_t \qquad \& \qquad C_t = K_t^{-1}$$

Also , from the inverse-Wishart component , we have

$$S_t = E_t - U_t^T.M_t \qquad \& \qquad n_t = a_t - d$$

respectively the point estimate of V and its degree of freedom . Then , we have established the equivalence between the Reference Analysis updating equations and the traditional standard updating equations for the common components multivariate DLM .

Comment : Reference form and the Jeffreys' prior

The reference form used in the Theorem 5.1 can be justified or interpreted in terms of the so called Jeffreys' priors . The Jeffreys' rule for multiparameter problems [see Box and Tiao(1973) , pg 54] can be stated as follows : ' The initial prior distribution for a set of parameters is taken to be proportional to the square root of the determinant of the information matrix '. For the prior distribution of (Θ_t, V) , we shall first of all assume that Θ_t and V are approximately independent so that , considering the standard (locally uniform) reference prior for Θ_t, $p(\Theta_t) \propto constant$, we have as joint reference prior , $p(\Theta_t, V) \propto p(V)$. Then , considering Jeffreys' rule for multiparameter problems , we have

$$p(\Theta_t, V) \propto |I(V)|^{\frac{1}{2}}$$

where I(V) is the Information matrix (minus the expected value of the second derivative of the log-likelihood) . Since $p(\Theta_t, V^{-1}) = p(\Theta_t, V).|\frac{\partial V}{\partial V^{-1}}|$, we have , using standard results and derivating the log-likelihood twice with respect to V^{-1} ,

$$|I(V)| = |I(V^{-1})|.|\frac{\partial V}{\partial V^{-1}}|^{-2} \propto |\frac{\partial V}{\partial V^{-1}}|^{-1}$$

Finally , it is necessary to proof that $|\frac{\partial V}{\partial V^{-1}}| = V|^{d+1}$ (see for instance Anderson(1984) , pg 601) , which gives the reference form used in the Theorem 5.1 ,

$$p(\Theta_t, V) \propto |\frac{\partial V}{\partial V^{-1}}|^{-\frac{1}{2}} = V|^{-\frac{1}{2}.(d+1)}$$

For further discussion about non-informative priors , an alternative reference (in a Kalman filter context) is Catlin(1989) , chapter 7 .

5.4.2 - Reference Analysis Implementation

Although the updating equations defined in the previous theorem , part A , are valid for all $t \geq 0$, for computational and interpretation reasons (avoidance of matrix inversions , easier interpretation ,etc) ,it is preferable to use the standard updating equations (Appendix - 5.2) for $t \geq t_P$, since both algorithms are algebraicly equivalent at these time points . For $0 \leq t < t_P$ however , where there is no such equivalence , we need to implement the reference analysis algorithm , and one major difficulty is the setup of the covariance matrices W_t . Since the posterior covariance matrices C_t do not exist for $t < t_P$ (K_t does not have full rank)

we can not apply the traditional discount techniques used in the implementation of standard conjugate prior analysis of DLM's .

The procedure used here is in line with the Pole and West(1987) method for the univariate case , and consists in assuming $W_t = 0$ for $t = 1,2,..,t_P$. This practical procedure has its rationale in the fact that is not possible to detect or estimate any changes in parameters during the first t_P observations , since we have only one observation for each parameter in Θ_t or V , and so , for convenience we set them to zero . The basic result necessary for the practical implementation of these ideas is given by the following theorem .

Theorem 5.2 : In the context of Reference Analysis for the common components multivariate DLM , suppose that G_t is non-singular and $W_t = 0$. Then , the prior and posterior distributions of Θ_t and V have the same forms as before (Theorem 5.1) , with the same initialization and observation updating equations but different time updating equations , as follows ,

i) Time Updating :

$$H_t = G_t^{T^{-1}}.K_{t-1}.G_t^{-1} \tag{5.27}$$

$$T_t = G_t^{T^{-1}}.U_{t-1} \tag{5.28}$$

$$L_t = E_{t-1} \tag{5.29}$$

ii) Observation Updating :

$$K_t = H_t + F_t.F_t^T \tag{5.30}$$

$$U_t = T_t + F_t.\underline{y}_t \tag{5.31}$$

$$E_t = L_t + \underline{y}_t.\underline{y}_t^T \tag{5.32}$$

$$a_t = a_{t-1} + 1 \tag{5.33}$$

iii) Initialization : $H_1 = 0$; $T_1 = 0$; $L_1 = 0$; $a_1 = 0$.

Proof : Suppose that the joint posterior distribution at time $t-1$ has the stated form as in the last theorem , given by

$$p(\Theta_{t-1},V/D_{t-1}) \propto |V|^{-\frac{1}{2}.(a_{t-1}+d+1)}.exp\{-\frac{1}{2}.trV^{-1}.(\Theta_{t-1}^T.K_{t-1}.\Theta_{t-1}-2.\Theta_{t-1}^T.U_{t-1}+E_{t-1})\}$$

Now , since $W_t = 0$, the system equation is $\Theta_t = G_t.\Theta_{t-1}$ which can be inverted (G_t is non-singular) so that $\Theta_{t-1} = G_t^{-1}.\Theta_t$, what is a linear transformation with constant Jacobian . Then , the joint prior distribution is immediately obtained as ,

$$p(\Theta_t, V/D_{t-1}) \propto |V|^{-\frac{1}{2}(a_{t-1}+d+1)} exp\{-\frac{1}{2} trV^{-1}(\Theta_t^T.G_t^{T^{-1}} K_{t-1} G_t^{-1}\Theta_t - 2\Theta_t^T G_t^{T^{-1}} U_{t-1} + E_{t-1})\}$$

$$\propto |V|^{-\frac{1}{2}(a_{t-1}+d+1)}.exp\{-\frac{1}{2}.trV^{-1}.(\Theta_t^T.H_t.\Theta_t - 2.\Theta_t^T.T_t + L_t)\}$$

where $H_t = G_t^{T^{-1}}.K_{t-1}.G_t^{-1}$; $T_t = G_t^{T^{-1}}.U_{t-1}$ & $L_t = E_{t-1}$.

Since the joint prior distribution and the likelihood are the same considered in the last theorem , the joint posterior distribution will be also the same , given by

$$p(\Theta_t, V/D_t) \propto |V|^{-\frac{1}{2}(a_t+d+1)}.exp\{-\frac{1}{2}.trV^{-1}.(\Theta_t^T.K_t.\Theta_t - 2.\Theta_t^T.U_t + E_t)\}$$

and the theorem is proved by induction .

Corollary : As a consequence of Theorems 5.1 and 5.2 , a practical Reference Analysis updating algorithm for the common components multivariate DLM is given as follows :

i) For $t = 0, 1, 2, .., t_P - 1$, where t_P is the minimum time such that the posterior distributions are proper ($t_P = p + \frac{1}{2}.d.(d+1)$ if there is no missing data at the begining , etc) , use the updating equations of Theorem 5.2 (initialization plus the cycle observation updating / time updating) .

ii) For $t \geq t_P$, use the standard matrix-normal / inverse-Wishart updating algorithm of Appendix 5.2 (equations 5.41-5.48a)

Comments :

i) In the particular case of dimension $d = 1$ (univariate case) , the results of Theorems 5.1 and 5.2 coincide with the Reference Analysis of univariate DLM's , as presented by Pole and West(1987) .

ii) In general , the assumptions of non-singularity of W_t and G_t present respectively in Theorems 5.1 and 5.2 , are satisfied for most practical situations . Even in rare cases such as when we have a moving-average component in the model , this difficulty can be avoided by considering an alternative parametrization .

5.4.3 - An Application with Real Data

Here we show in practice one application with real data using the algorithm for Reference Analysis of multivariate DLM's with common components presented in the last sub-section (Corollary of Theorems 5.1 and 5.2) .

The data set to be analysed consists of a 3-dimensional vector time series of (monthly) Industrial Production Indexes-IPI in Spain , since January of 1981 to August of 1988 . The three marginal component time series consist of the following categories or types of industrial production indexes : Consumer IPI , Investment IPI and Intermediate IPI . This vector valued time series is shown graphically in figure 5.1 , at the end of this chapter , where we can observe some long-term trend behaviour and a seasonal pattern in the data .

The presence of a trend or growth component in this kind of data is not surprising since Spain is one of the fast growing economies in the European comunity , although the presence of significant positive levels of growth in the investment IPI series has been observed only in more recent years , after the middle of 1985 (see fig. 5.1) . In particular , in relation to the seasonal behavior , we can notice from that diagram , a very sharp fall in the industrial production indices of the three series in August of each year . The reason for such an abrupt change in the industrial production index at this particular month is because the holiday period for all industrial workers in Spain is in August , when the production is reduced to minimum levels .

As a result , to express this sort of 'holiday-effect' seasonality we need six harmonics (the first or 12-period harmonic , the sixth or 2-period harmonic corresponding to the Nyquist frequency and all the intermmediate ones) since , apart from trend or other components , the levels of IPI in July of each year , for all the three series , are recovered again two periods (months) later in September , which suggests the presence of six harmonics . The presence of all these 6 harmonics is also confirmed through the use of Periodograms for all the three component series , so the overall seasonal structure is described by the first six harmonics .

The trend component is considered as a dynamic linear growth model (level and slope or growth rate parameter for each marginal time series) , and the model specification is then complete . One important aspect to be mentioned here related with model specification is that the assumption that all the six harmonics are present in all marginal series (common seasonal structure) is a very particular one , valid approximately for the specific data set in question , but not as a general rule . More general models may be necessary and this is discussed in the

next chapter .

The residuals from the one-step-ahead forecasts considering the present model (linear trend + six harmonics) , using the Reference Analysis algorithm for common components multivariate DLM's of last sub-section is shown in figure 5.2 . As we can notice , the number of observations per series necessary in order to have proper distributions is 15 (11 for the six harmonics + 2 for the linear trend + 1 noise variance + 1 correlation) , and after that all distributions are proper , giving consequently , among others , one-step-ahead forecasts and its residuals (the codes 1 , 2 and 3 are used to refer respectively to Consumer IPI , Investment IPI and Intermediate IPI) . Also , and the more important aspect we can observe in fig 5.2 , is the fact that the residuals are centered around zero and reasonably *uncorrelated only for the* series 1 and 3 , respectively Consumer IPI and Intermediate IPI . For the series of Investment IPI however , after middle of 1985 approximately , the residuals are not more centered around zero and are highly positively correlated , resulting in a significant under forecasting . In fact , this is the result of a major change in the trend pattern for the series of Investment IPI around middle of 1985 , as we can see in fig 5.1 . Consequently , some sort of intervention in the estimation process for this series around middle of 1985 is necessary in order to restore the previous forecasting performance , but this is not possible within the common component structure as mentioned before .

Also , the correlation structure among the series is given as a result from the mentioned algorithm , which is shown in figure 5.3 . As we can see from that picture , the correlation between the comsumer and investment industrial production indexes is about 0.5 approximatelly , and the same is valid for the correlation between the investment and the intermediate IPI . The correlation between the comsumer and the intermediate IPI however is not so stable in the begining , but after the first half of the series it reaches values about 0.6 . In fact we were not expecting very high correlations among the series (or , among the correspondent error terms) since the trend component is not very expressive .

Finally , in order to complement the present discussion about model identification , it is interesting to observe two aspects . First , as mentioned before , the assumption that all the three series have a six harmonic structure (common component hypothesis) is only one approximation . From the Periodogram graphics shown in figs. 5.4 to 5.6 , we can see for instance that the importance of the first harmonic (in terms of percentage of variability explained in

the data) is very different in each series . If we take into account some considerations about model parsimony , would be possible to drop the first harmonic in one series or two perhaps , but not in all of then , and the model would not be a common component one .

Second , as an alternative (more parsimonious) model formulation , it is interesting to mention that three harmonics (the first , the fourth and the sixth) involving a total of 5 parameters per series could be substituted by only one regressor coefficient per series , where the regressor is a dummy variable assuming the value 1 corresponding to every August month and zero otherwise .

Then , this real example has suggested us that the common component model is only one first approximation presenting serious drawbacks . If we want a more parsimonious seasonal representation or more proper forecasts for the series of investments , then we need go beyond the common component structure .

Appendix 5.1 : The Matrix Inversion Lemma and the Reference Analysis of D.L.M.'s

Lemma (Matrix Inversion) : Given the matrices R , F and V , with dimensions respectively pxp , pxn and nxn , then , the following algebraic identity holds , supposing that the various inverses exist

$$(R^{-1} + F.V^{-1}.F^T)^{-1} = R - R.F.(F^T.R.F + V)^{-1}.F^T.R$$

Proof : See , for instance , Anderson & Moore(1979) , pg 138 or Quintana(1987) , pg 51 .

Theorem (Reference Analysis) : Consider the multivariate DLM defined by the equations (5.1) and (5.2) where V_t is known , G_t is of full rank and W_t is non-singular for all t . If we assume that the initial prior information for the process parameter $\underline{\theta}_t$ is represented by the reference form

$$p(\underline{\theta}_1 / D_0) \qquad \propto \qquad \text{constant}$$

then , the prior and posterior distributions of $\underline{\theta}_t$ at time t = 1,2,.. have the following (possibly improper) form

$$p(\underline{\theta}_t / D_{t-1}) \propto exp\{-\frac{1}{2}(\underline{\theta}_t^T.H_t.\underline{\theta}_t - 2.\underline{\theta}_t^T.\underline{h}_t)\}$$

$$p(\underline{\theta}_t / D_t) \propto exp\{-\frac{1}{2}(\underline{\theta}_t^T.K_t.\underline{\theta}_t - 2.\underline{\theta}_t^T.\underline{k}_t)\} \qquad \text{where}$$

$$H_t = W_t^{-1} - W_t^{-1}.G_t.P_t^{-1}.G_t^T.W_t^{-1} \tag{5.34}$$

$$P_t = G_t^T.W_t^{-1}.G_t + K_{t-1} \tag{5.34a}$$

$$\underline{h}_t = W_t^{-1}.G_t.P_t^{-1}.\underline{k}_{t-1} \tag{5.35}$$

$$K_t = H_t + F_t.V_t^{-1}.F_t^T \tag{5.36}$$

$$\underline{k}_t = \underline{h}_t + F_t.V_t^{-1}.\underline{y}_t \tag{5.37}$$

with initial values $H_1 = 0$ and $\underline{h}_1 = 0$.

Proof : The proof proceeds by induction on t . The likelihood for $\underline{\theta}_t$ from an observation $\underline{y}_t$ is given by

$$l(\underline{\theta}_t/\underline{y}_t) \propto exp\{-\frac{1}{2}(\underline{\theta}_t^T.F_t.V_t^{-1}.F_t^T - 2.\underline{\theta}_t^T.F_t.V_t^{-1}.\underline{y}_t)\}$$

Supposing that $p(\underline{\theta}_t/D_{t-1})$ is given as stated , then

$$\begin{aligned} p(\underline{\theta}_t/D_t) &\propto p(\underline{\theta}_t/D_{t-1}).l(\underline{\theta}_t/\underline{y}_t) \\ &\propto exp\{-\frac{1}{2}(\underline{\theta}_t^T.H_t.\underline{\theta}_t - 2.\underline{\theta}_t^T.\underline{h}_t + \underline{\theta}_t^T.F_t.V_t^{-1}.F_t^T.\underline{\theta}_t - 2.\underline{\theta}_t^T.F_t.V_t^{-1}.\underline{y}_t)\} \\ &\propto exp\{-\frac{1}{2}(\underline{\theta}_t^T.K_t..\underline{\theta}_t - 2.\underline{\theta}_t^T.\underline{k}_t)\} \end{aligned}$$

and the posterior for $\underline{\theta}_t$ is also of the stated form . Now , the prior for the next time $t+1$ will be

$$p(\underline{\theta}_{t+1}/D_t) = \int p(\underline{\theta}_{t+1}, \underline{\theta}_t/D_t).d\underline{\theta}_t = \int p(\underline{\theta}_{t+1}/\underline{\theta}_t, D_t).p(\underline{\theta}_t/D_t).d\underline{\theta}_t$$

From the system equation (5.2) , the first term in this integral is a normal distribution with mean $G_{t+1}.\underline{\theta}_t$ and covariance matrix W_{t+1} , which gives ,

$$\begin{aligned} p(\underline{\theta}_{t+1}/D_t) &\propto \int exp\{-\frac{1}{2}[(\underline{\theta}_{t+1} - G_{t+1}.\underline{\theta}_t)^T.W_{t+1}^{-1}.(\underline{\theta}_{t+1} - G_{t+1}.\underline{\theta}_t) + \underline{\theta}_t^T.K_t.\underline{\theta}_t - 2.\underline{\theta}_t.\underline{k}_t]\}.d\underline{\theta}_t \\ &\propto \int exp\{-\frac{1}{2}[(\underline{\theta}_{t+1} - \underline{\alpha}_t)^T.P_t.(\underline{\theta}_{t+1} - \underline{\alpha}_t) + \gamma_t]\}.d\underline{\theta}_t \qquad \text{where} \\ P_t &= K_t + G_{t+1}.W_{t+1}^{-1}.G_{t+1} \\ \underline{\alpha}_t &= P_t^{-1}.(\underline{k}_t + G_{t+1}^T.W_{t+1}^{-1}.\underline{\theta}_{t+1}) \\ \gamma_t &= \underline{\theta}_{t+1}^T.W_{t+1}^{-1}.\underline{\theta}_{t+1} - \underline{\alpha}_t^T.P_t.\underline{\alpha}_t \end{aligned}$$

(Note that P_t is non-singular since K_t is symmetric positive semi-definite , W_{t+1} is non-singular and G_{t+1} is of full rank) . Then , by standard normal theory , we have

$$p(\underline{\theta}_{t+1}/D_t) \propto exp\{-\frac{1}{2}\gamma_t\}$$

where , expanding γ_t , we have :

$$\gamma_t = \underline{\theta}_{t+1}^T . W_{t+1}^{-1} . \underline{\theta}_{t+1} - (\underline{k}_t + G_{t+1}^T . W_{t+1}^{-1} . \underline{\theta}_{t+1})^T . P_t^{-1} . (\underline{k}_t + G_{t+1}^T . W_{t+1}^{-1} . \underline{\theta}_{t+1})$$

$$= \underline{\theta}_{t+1}^T . H_{t+1} . \underline{\theta}_{t+1} - 2 . \underline{\theta}_{t+1}^T . \underline{h}_{t+1} + const., \qquad \text{with}$$

$$H_{t+1} = W_{t+1}^{-1} - W_{t+1}^{-1} . G_{t+1} . P_t^{-1} . G_{t+1}^T . W_{t+1}^{-1}$$

$$\underline{h}_{t+1} = W_{t+1}^{-1} . G_{t+1} . P_t^{-1} . \underline{k}_t$$

Thus , we have shown that if the prior at time t has the stated form then the prior at time t+1 also has the stated form , and also that the posteriors are of the stated form . It remains to show that the initial prior distribution satisfies the theorem . This is true because $H_1 = 0, \underline{h}_1 = 0$ imply $p(\underline{\theta}_1 / D_0) \propto exp\{-\frac{1}{2}.0\}$, or $p(\underline{\theta}_1 / D_0) \propto$ constant , which is exactly the reference form , and the theorem is proved .

Comment : The equations (5.36)-(5.37) of the Reference Analysis (correspondent to the posterior moments in the standard updating) coincide with the corresponding equations of the Information or Inverse Covariance Filter - Anderson & Moore (1979) , pg 139-140 . The other equations (5.34 - 5.35) do not coincide with the filter equations , although they are algebraically equivalent to them . Also , the equations (5.34) - (5.35) coincide with the equations presented by Pole & West (1987) , pg 7 , and the equations (5.36)-(5.37) are a natural multivariate extension of those equations since in the unidimensional case , we can rewrite (5.36)-(5.37) as

$$K_t = H_t + F_t . F_t^T . V_t^{-1} \qquad ; \qquad \underline{k}_t = \underline{h}_t + F_t . y_t . V_t^{-1}$$

where V_t^{-1} and y_t are now scalars .

Appendix 5.2 : Basic Notation and the Matrix Normal /Inverse Wishart Updating

Notation , Definitions and Properties :

a) vec A denotes the usual column-vectorization of a matrix A , i.e. , if A = ($\underline{a}_1, ..., \underline{a}_m$) , then , vec A = ($\underline{a}_1^T, ..., \underline{a}_m^T$) and some of the more useful properties include the following :

$$i) \quad vec(A + B) = vecA + vecB$$

$$ii) \quad vecA.C.B = (B^T \otimes A).vecC$$

b) A ⊗ B denotes the Kronecker direct product between the matrices A and B , i.e. ,

$$A \otimes B = \begin{pmatrix} a_{11}.B & \dots & a_{1n}.B \\ \vdots & . & \vdots \\ a_{m1}.B & \dots & a_{mn}.B \end{pmatrix}$$

and some basic properties include the following :

$$i) \quad I \otimes A = diag(A, .., A)$$

$$ii) \quad (A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$

$$iii) \quad (A \otimes B).(C \otimes D) = A.C \otimes B.D$$

c) The random matrix Θ is said to have a Matrix Normal Distribution ,

$$\Theta \sim N(M, C, \Sigma) \quad i.f.f. \quad vec\,\Theta \sim N(vecM, \Sigma \otimes C)$$

what means that if the matrices C and Σ are both positive definite , the probability density of Θ is proportional to ,

$$exp\{-\frac{1}{2}tr(\Theta - M).C^{-1}.(\Theta - M).\Sigma^{-1}\}$$

As a consequence , the linear transformation $A.\Theta.B + D$ is also a matrix normal distribution , with mean A.M.B + C and covariance matrices (left and right) given by $A.C.A^T$ and $B^T.\Sigma.B$.

Also , the matrices Θ and Σ are said to have a joint Matrix Normal / Inverse-Wishart distribution ,

$$\begin{pmatrix} \Theta \\ \Sigma \end{pmatrix} \sim NW^{-1}(M, C, S, d)$$

if and only if $\Theta/\Sigma \sim N(M, C, \Sigma)$ and $\Sigma \sim W^{-1}(S, d)$.

The Matrix Normal / Inverse-Wishart Model :

The common components model of section 5.3 can be equivalently formulated in a more compact matrix-normal distribution notation writting the vector $\underline{\theta}_t$ as a matrix Θ_t where each column represents the process parameters associated with each univariate marginal D.L.M. ($\underline{\theta}_t = vec\Theta_t$) as follows:

$$\text{i) observation equation:} \quad \underline{y}_t^T = F_t^T.\Theta_t + \underline{v}_t^T, \qquad \underline{v}_t^T \sim N(\underline{0}, V) \tag{5.38}$$

$$\text{ii) system equation:} \quad \Theta_t = G_t.\Theta_{t-1} + U_t, \quad U_t \sim N(O, W_t, V) \tag{5.39}$$

iii) prior information: $(\Theta_{t-1}/V, D_{t-1}) \sim N(M_{t-1}, C_{t-1}, V)$ (5.40)

$$(V/D_{t-1}) \sim W^{-1}(S_{t-1}, n_{t-1}) \tag{5.40a}$$

Obs: This model formulation was presented originally by Quintana(1985) , where further details and references can be found.

The equivalence between the matrix-normal formulation of equations (5.38)- (5.40a) and the standard D.L.M. formulation given by the equations (5.14)- (5.16a) can be easily verified: from (5.16),

$$(vec\Theta_{t-1}/V, D_{t-1}) \sim N(vecM_{t-1}, V \otimes C_{t-1})$$

$$\leftrightarrow (\Theta_{t-1}/V, D_{t-1}) \sim N(M_{t-1}, C_{t-1}, V)$$

, what is (5.40) ; in a similar way, the other two equivalences can be verified.

Finally, the Updating Relations for $\underline{\theta}_t$ and V_t can be obtained considering the standard multivariate-normal/inverse-Wishart model - eq.(5.14)-(5.16a) or, alternatively, using the more compact matrix-normal notation, we get the following updating equations:

$$M_t = M_t^* + A_t.E_t \tag{5.41}$$

$$C_t = R_t - A_t.Q_t.A_t^T \tag{5.42}$$

$$M_t^* = G_t.M_{t-1} \tag{5.43}$$

$$R_t = G_t.C_{t-1}.G_t^T + W \tag{5.44}$$

$$A_t = R_t.F_t^T.Q_t^{-1} \tag{5.45}$$

$$e_t = \underline{y}_t - F_t^T.M_t^* \tag{5.46}$$

$$Q_t = F_t^T.R_t.F_t + 1 \tag{5.47}$$

and also $(V/D_t) \sim W^{-1}(d_t, n_t)$ where :

$$d_t = d_{t-1} + e_t.e_t^T.Q_t^{-1} \tag{5.48}$$

$$n_t = n_{t-1} + 1 \tag{5.48a}$$

FIG 5.1 INDUSTRIAL PRODUCTION INDEXES IN SPAIN

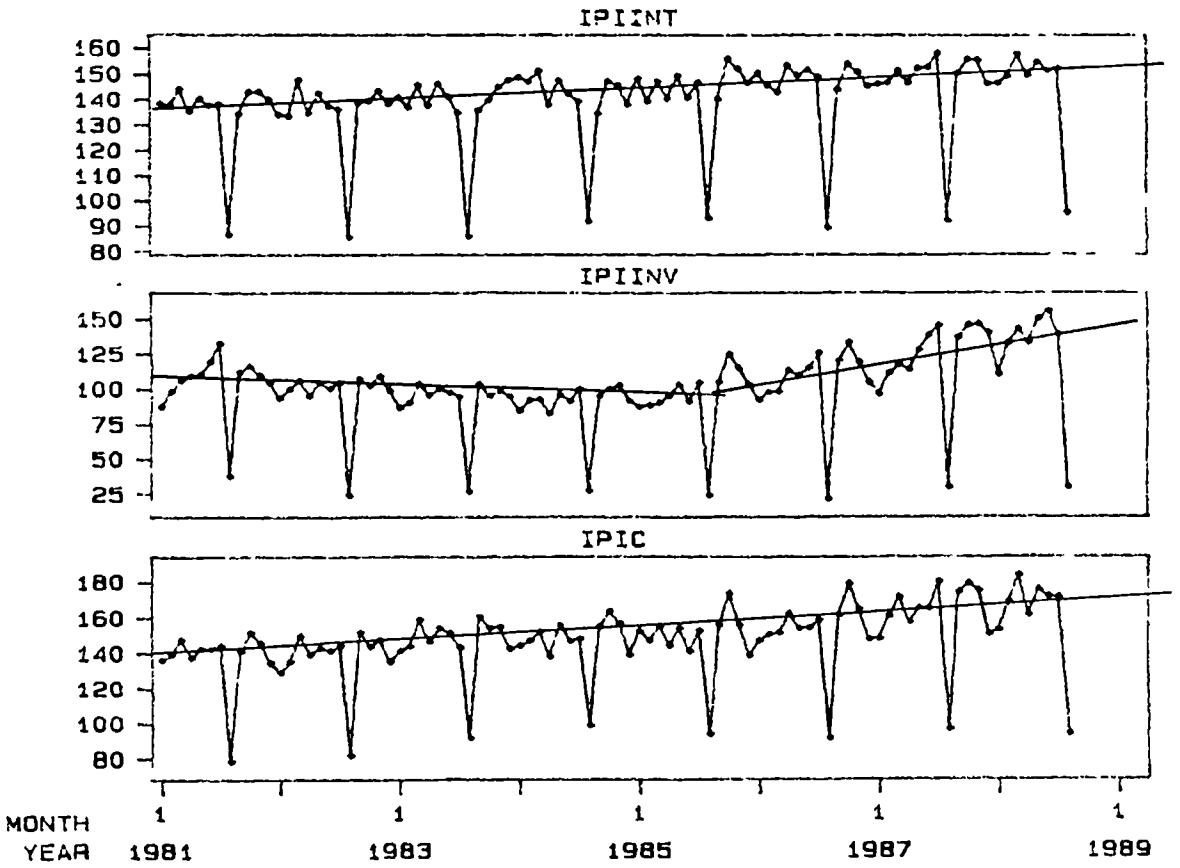


FIG. 5.2 ONE STEP AHEAD FORECASTING RESIDUALS

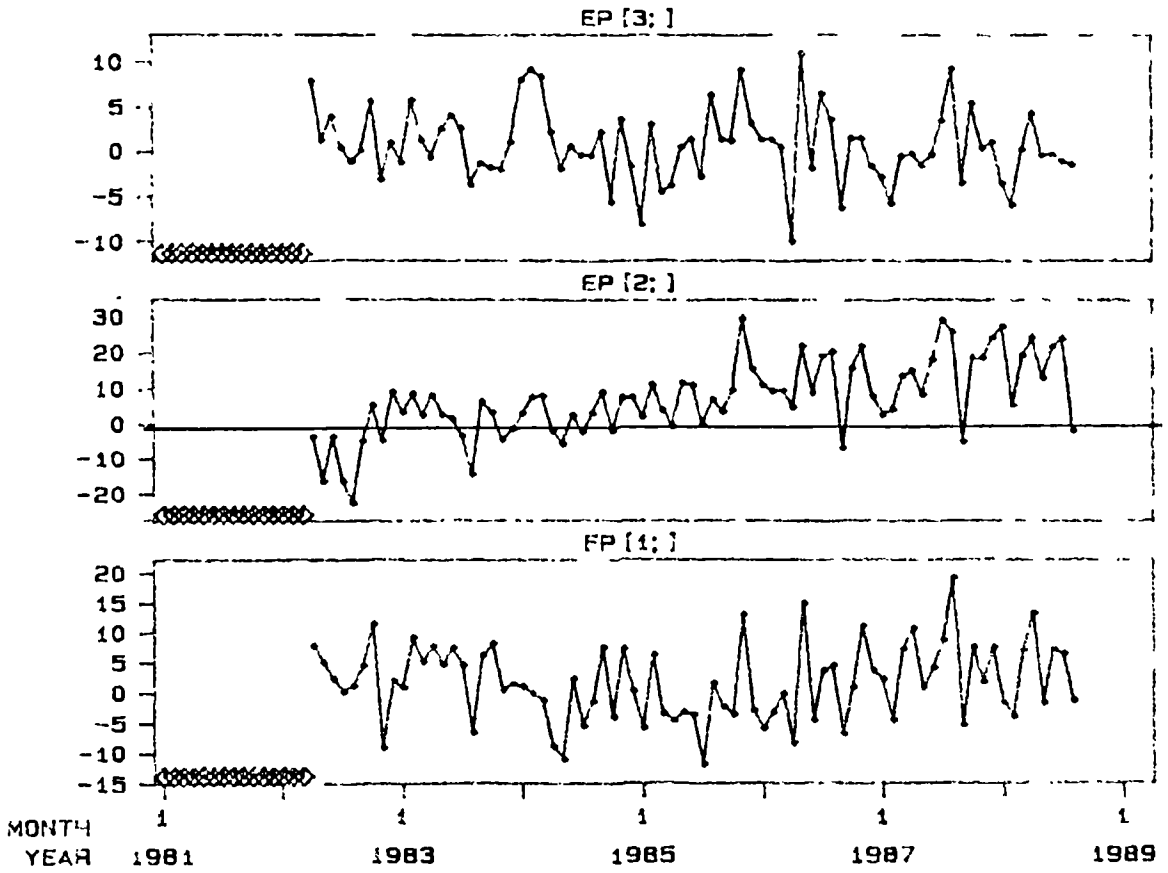


FIG. 5.3 CORRELATION BETWEEN THE COMPONENT SERIES

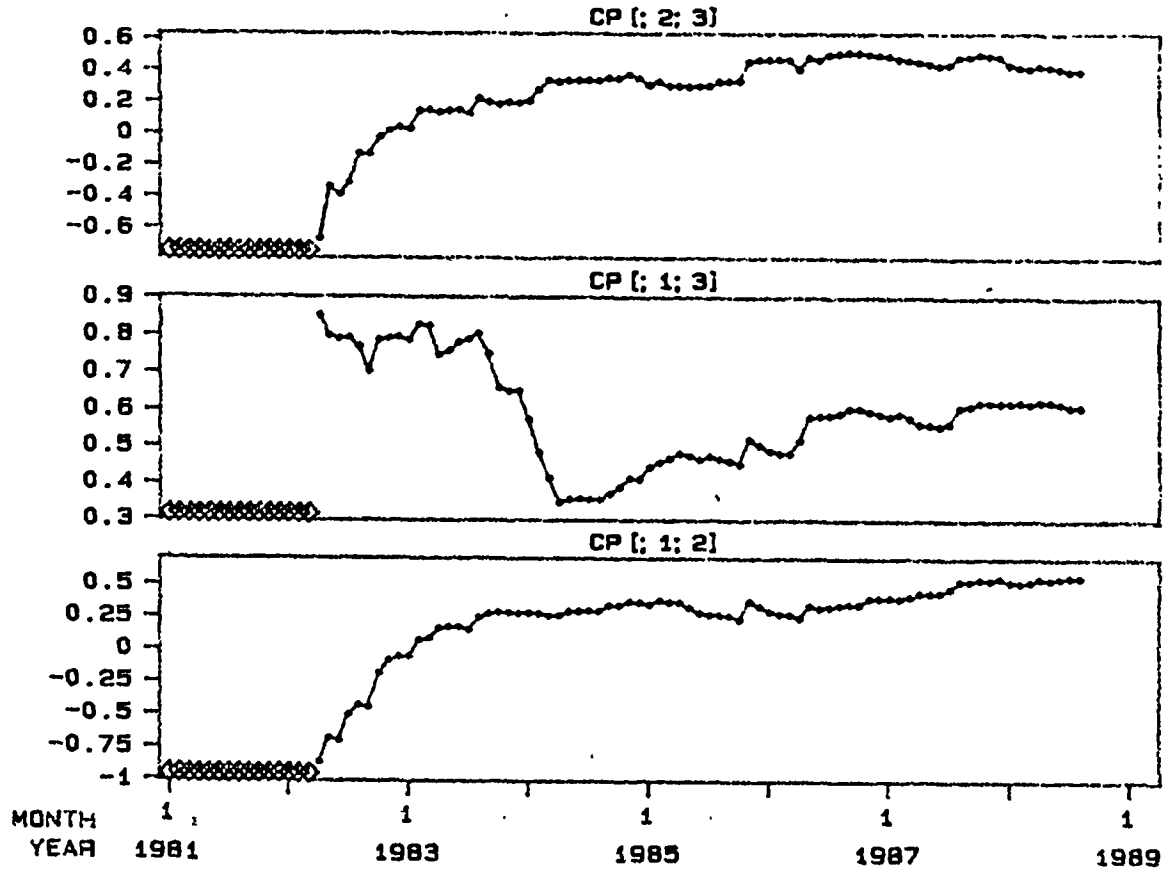


FIG 5.4 -- PERIODOGRAM -- CONSUMER IPI. SERIES

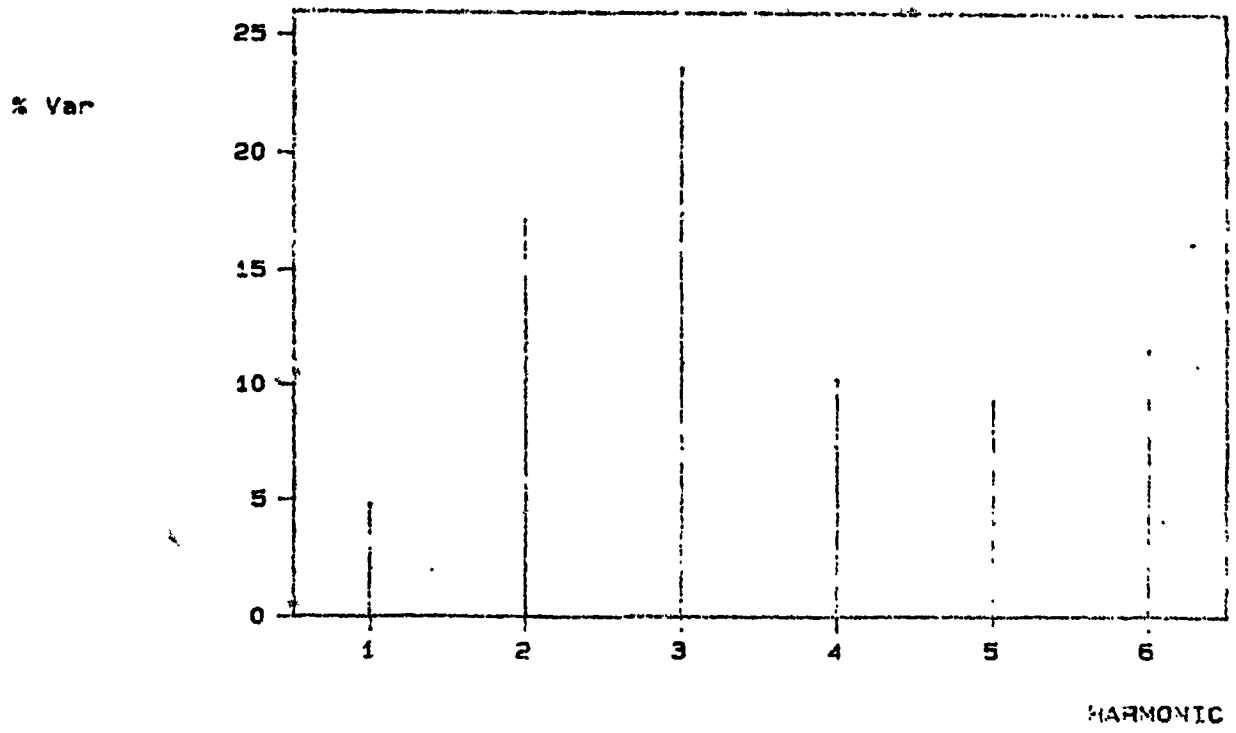


FIG 5.5 -- PERIODOGRAM -- INVERSION IPI SERIES

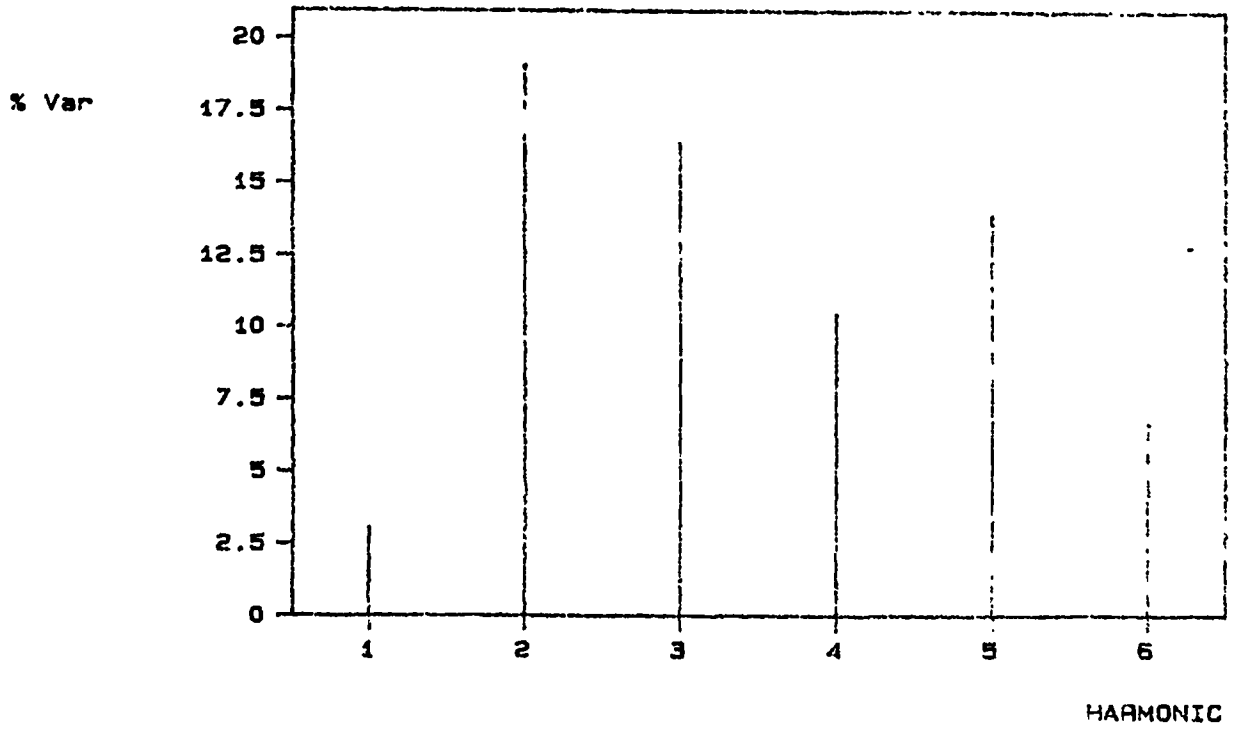
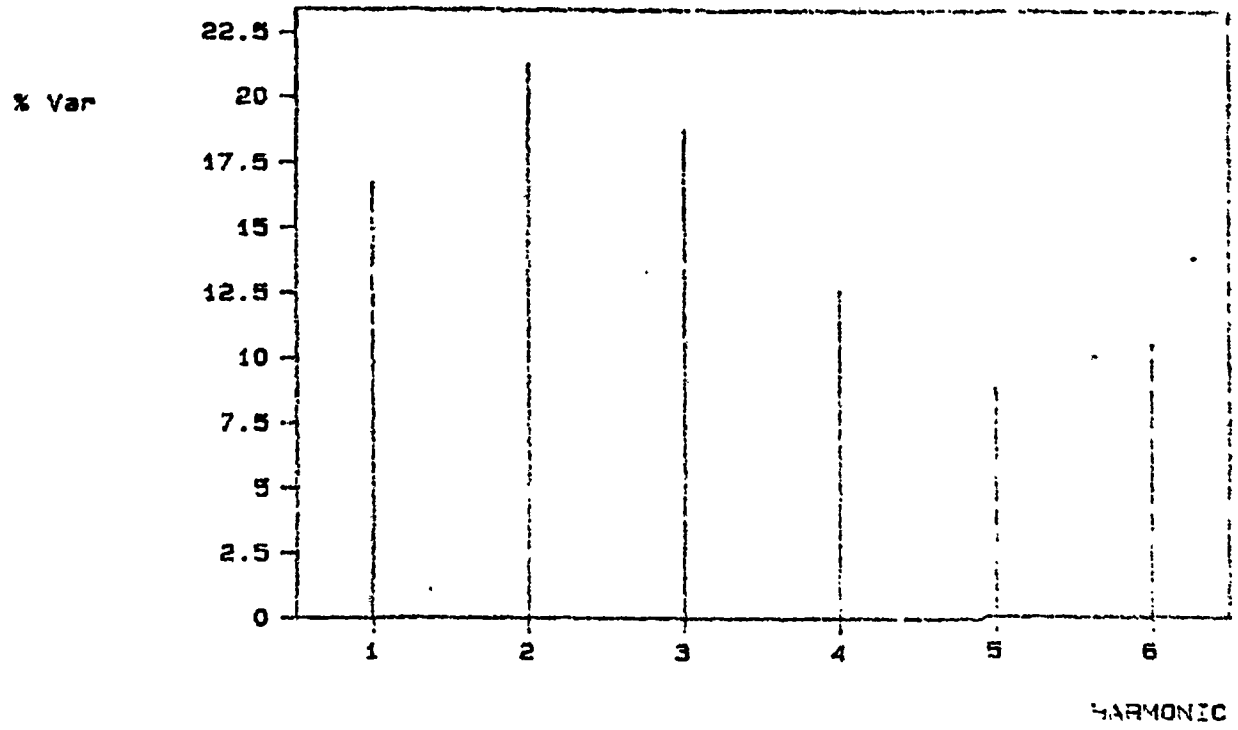


FIG 5.6 -- PERIODOGRAM -- INTERMEDIATE IPI SERIES



CHAPTER 6

A GENERAL METHODOLOGY FOR D.L.M. ANALYSIS

6.1 Introduction

We introduce here a general methodology for analysis of multivariate D.L.M.'s where the vector process parameter $\underline{\theta}_t$ and the observational noise variance-covariance matrix V are both estimated on line in a fully joint fashion, without the limitations present in the previous model of the last chapter (the design of the matrices F and G does not need to have common blocks only , and so on) . This is possible considering a multivariate extension of some ideas used in the univariate case. As we have seen from section 5.3.1 (unidimensional case) one of the key ideas abo t estimating $\underline{\theta}_t$ when V is unknown is the use of a factor to correct the standard structure for the posterior variance C_t . This factor (see equation 5.13b) is basically V or its last estimate S_t with reference to the previous estimate S_{t-1} or any other reference as for instance an initial reference S_0. In this way we are introducing the uncertainty about V in the estimation equations for $\underline{\theta}_t$, that gives a fully joint estimation procedure .

In the multivariate case , V is not a scalar and generally C_t and S_t do not have the same dimensions so that this principle cannot be applied in a straight forward way. But this idea (also present in the common component model of section 5.3.2 where all covariance matrices are self-scaled by V in a Kronecker product fashion) is very attractive and should be pursued.

This dimension problem however can be avoided by reparametrization from the process parameter $\underline{\theta}_t$ to the mean response parameter $\underline{\mu}_t = F_t.\underline{\theta}_t$, since the variance of $\underline{\mu}_t$ has the same dimension dxd as V and can be scaled by S_t.

A general and natural variance structure considers symmetry , where all covariance matrices are symmetrically pre and post multiplied by $V^{\frac{1}{2}}$ in a simple multidimensional extension of the univariate formulation . In fact , this variance structure is compatible not only with the univariate case , but also with the common components model formulation , and in a more general multivariate DLM framework , it provides the structure for a joint sequential estimation (approximated conjugate prior analysis) of $\underline{\theta}_t$ and V .

When we introduce this key feature in the model structure , the adaptive coefficient and the posterior mean for $\underline{\theta}_t$ will depend on V (and consequently on the inter-series correlations) , so giving a fully multivariate model without the constraints mentioned before.

It is important to remember however that the D.L.M. model has its dynamic evolution expressed in terms of the process parameter $\underline{\theta}_t$ and consequently we will need both parametrizations: $\underline{\theta}_t$ in the time updating and $\underline{\mu}_t$ in the observation updating, where the basic model structure is defined.

As this sort of variance-scaling operation is specific to the $\underline{\mu}_t$ parametrization, we should apply the inverse operation before the inverse reparametrization from $\underline{\mu}_t$ to $\underline{\theta}_t$ is carried out again.

The posterior moments for $\underline{\mu}_t$ are transformed back into the corresponding moments for $\underline{\theta}_t$ exploiting the Conditional Independence of $\underline{\theta}_t$ and $\underline{y}_t$ given $\underline{\mu}_t$. The necessary results about Conditional Independence are presented in Appendix 6.1A and the full model formulation and analysis is presented in the next section as well as the corresponding updating equations and some basic properties . In the following section of this chapter , an alternative representation for the same model is proposed with different characteristics, and implemented in two alternative versions. Finally , in the last section , one example of application using real data is given .

6.2 A General Multivariate D.L.M.

Definition: A scaled version of the general multivariate D.L.M. model for a vector of observations $\underline{y}_t$ of dimension d made at intervals , at times t=1,2,...is defined by equations (5.1) and (5.2) with the following distributional assumptions:

$$\text{(i) Likelihood function : } \quad (\underline{y}_t/\underline{\mu}_t, V = S^2) \sim N(\underline{\mu}_t; S^2) \tag{6.1}$$

$$\text{(ii) Approximate Prior Distribution : } \quad (\underline{\mu}_t/D_{t-1}, V) \sim N(\underline{f}_t; R_t^*) \tag{6.2}$$

$$(V^{-1}/D_{t-1}) \sim W(U_{t-1}^{-2}; n_{t-1}) \tag{6.3}$$

where:

$\underline{\mu}_t = F_t.\underline{\theta}_t$ with F_t & $\underline{\theta}_t$ as defined before is the mean response parameter.

$\underline{f}_t$ is the prior mean for $\underline{\mu}_t$ given V

$R_t^* = S.\Sigma_t.S$ is the prior variance-covariance matrix for $\underline{\mu}_t$ given V , where Σ_t is given by $\Sigma_t = Var[\underline{\mu}_t/D_{t-1}, V = I]$

U_{t-1}^{-2} is the prior precision matrix, and n_{t-1} is the prior d.f.

Comment :

It is important to notice that any symmetric positive definite matrix V can be expressed as the square of a matrix S as in (6.1) . Although S is not unique , expedient choices of its form are given in Appendix 6.2b .

6.2.1 Model Analysis:

Considering the assumptions (6.1)-(6.3) , the Bayesian analysis of the scaled multivariate D.L.M. gives:

(i) Predictive distribution:

$$(\underline{y}_t/D_{t-1}, S^2) \sim N(\underline{f}_t; Q_t^*) \tag{6.4}$$

where $Q_t^* = S.Q_t.S$, with $Q_t = \Sigma_t + I = q_t^{-2}$ say. This predictive distribution for $\underline{y}_t$ (note that $\underline{y}_t = \underline{\mu}_t + \underline{v}_t$) is obtained immediately from (6.1)-(6.2) since its mean vector $\underline{f}_t$ coincide with the mean of $\underline{\mu}_t$ given in (6.2) and its variance-covariance matrix Q_t^* is the sum of the two covariance matrices given in (6.1) and (6.2) .

(ii) The adaptive coeficient (regression of $\underline{\mu}_t$ on $\underline{y}_t$ given D_{t-1} & S^2) is defined by:

$$\begin{aligned} A^* &= Cov(\underline{\mu}_t, \underline{y}_t).[Var(\underline{y}_t)]^{-1} \quad \text{given } D_{t-1} \,\&\, S^2, \quad \text{or} \\ A^* &= Var(\underline{\mu}_t/D_{t-1}, S^2).[Var(\underline{y}_t/D_{t-1}, S^2)]^{-1} = S.\Sigma.S.(S.Q.S)^{-1} \\ A^* &= S.A.S^{-1} \quad \text{where} \quad A = \Sigma.Q^{-1} \end{aligned} \tag{6.5}$$

(iii) The Sequential Approximation of the Posterior Distribution :

$$(\underline{\mu}_t/D_t, S^2) \sim N[\underline{m}_t = \underline{f}_t + A^*.(\underline{y}_t - \underline{f}_t)\,;\, C_t = S.(I - A).\Sigma_t.S\,] \tag{6.6}$$

The expressions in (6.6) are easily verified from (6.4) and (6.5) . Given the adaptive coefficient matrix A^* , the expression for the mean in (6.6) follows immediately as in usual posterior mean updating . In the same way , the variance matrix expression in (6.6) can be obtained from our usual posterior variance updating , which using the present notation is $C_t = R_t^* - A_t^*.Q_t^*.A_t^{*T}$. Substituting R_t^* , A_t^* and Q_t^* by their respective expressions , we get the desired result .

Also , defining $H^T = S^{-1}.q.S$ or $H_t^T = S^{-1}.Q_t^{-\frac{1}{2}}.S$, we have , using (6.4) , that

$$(H^T.\underline{e}_t/D_{t-1}, S^2) \sim N(\underline{0}; V) \tag{6.7}$$

And , using (6.3) , by Bayes rule we get

$$(V^{-1}/D_t) \sim W(\frac{n_t.\tau + H^T.\underline{e}_t.\underline{e}_t^T.H}{n_t+1}; n_t+1) \qquad (6.8)$$

where τ is the precision matrix given by (6.3).

6.2.2 - Posterior approximation & related aspects :

The first point that is important to stress here is that a precise formulation of DLM's in closed form (exact conjugate prior analysis) is possible only for the scalar (univariate) and the common components model , but not for the general case .

For these two special cases , the adaptive coefficients matrix A^* does not depend on V (in fact , $A^* = A_t = R_t\ F.Q_t^{-1}$ for the univariate case and $A^* = \mathbf{A}_t = I \otimes A_t$ for the common components case) and consequently there is no approximation involved .

But in the general case , since $V = S^2$, we find that the adaptive coefficients matrix A^* is of the form $A^* = S.A.S^{-1}$ (and also $H^T = S^{-1}.q.S$) and some sort of approximation is necessary in order to keep a tractable close form of analysis .

A natural approach to this problem is to consider that in the locality of our best estimate of V , namely $V_{t-1} = S_{t-1}^2$, we have that the adaptive coefficient matrix A^* is locally approximated by $A^* \simeq S_{t-1}.A.S_{t-1}^{-1}$ and also similarly $H^T \simeq S_{t-1}^{-1}.q.S_{t-1}$.

The main practical and theoretical reasons that support the use of the approximations considered as well as the whole methodology are :

i) It is an effective procedure that enables a tractable closed form of analysis at a very reasonable computational cost (the whole algorithm is presented in the next sub-section with full implementation details)

ii) The proposed general formulation coincide with exact conjugate prior analysis results in the special cases of common components and univariate models , which is an excellent theoretical property (this one and other properties are presented with the corresponding proofs in section 6.2.4 of this chapter) .

iii) Even in the general case , the approximations involved can become exact or almost exact if all estimated elements in V have an error (difference from the unknown true variance or covariance values) proportional or almost proportional to a same constant . For a study of the sensitivity of $S.A.S^{-1}$ in relation to perturbations in S , see Appendix 6.1b.

iv) As the number of degrees of freedom increases , the model analysis for the general case became exact , since U_{t-1} converge in probability to S and we approach the case of a multivariate DLM where V is known , which is presented in section 5.2.2 . Then , even when the linear or proportional assumption is not verified , the approximations are still justified since in general U approaches S very quickly .

v) Finally , it is important to mention that the learning process for V is a key feature of the present method and in general , as shown in the next chapter , it is faster and better than with other approximation procedures for multivariate DLM analysis .

6.2.3 Updating Equations

In accordance with the general concepts introduced in 6.1 and the formal model definition and analysis of last sub-section, the updating equations for the scaled multivariate D.L.M. model will be given by the following steps:

notation: $(\underline{\theta}_{t-1}/D_{t-1}, V = I) \sim N(\underline{m}^1_{t-1}, C^1_{t-1})$

step 1 - time updating: $(\underline{\theta}_t/D_{t-1}, V = I) \sim N(\underline{a}_t, R_t)$ where:

$$\underline{a}_t = G.\underline{m}^1_{t-1} \tag{6.9}$$

$$R_t = G.C^1_{t-1}.G^T + W \tag{6.10}$$

where W is specified through a given vector of discount factors $\underline{b}$. There is no time updating for V - it is supposed to be constant, and the time evolution for $\underline{\theta}_t$ is the same as in the standard multivariate D.L.M. model of last chapter .

step 2 - reparametrization and scaling: $(\underline{\mu}_t/D_{t-1}, V) \sim N(\underline{f}_t, R^*_t)$ where:

$$\underline{f}_t = F^T_t.\underline{a}_t \tag{6.11}$$

$$R^*_t = S_c.\Sigma_t.S^T_c \tag{6.12}$$

$$\Sigma_t = F^T_t.R_t.F_t \tag{6.12a}$$

where initially, the scaling matrix S_c is set up as an identity and updated in step 3.

step 3 - observation updating. a) posterior moments for $\underline{\mu}_t$: $(\underline{\mu}_t/D_t, V) \sim N(\underline{m}_t, C_t)$ where:

$$\underline{m}_t = \underline{f}_t + A^*.(\underline{y}_t - \underline{f}_t) \tag{6.13}$$

$$C_t = R_t^* - A_t^*.Q_t^*.A_t^{*T} \tag{6.14}$$

$$A_t^* = R_t^*.Q_t^{*-1} \tag{6.14a}$$

$$Q_t^* = R_t^* + V \tag{6.14b}$$

b) posterior moments for V : $(V/D_t) \sim W^{-1}(d_t, n_t)$ where:

$$d_t = n_t.V_t = (d_{t-1} + h_t.h_t^T) \tag{6.15}$$

$$n_t = n_{t-1} + 1 \tag{6.15a}$$

$$h_t = S_{t-1}.[(Q_t^*)^{\frac{1}{2}}]^{-1}.\underline{e}_t \tag{6.15b}$$

$$S_t = (V_t)^{\frac{1}{2}} \tag{6.16}$$

$$S_c = S_t.S_0 \tag{6.16a}$$

where S_0 is a reference matr x set up initially such that the initial scale factor S_c is the identity matrix.

step 4 - Inverse reparametrization and scaling $(\underline{\theta}_t/D_t, V = I) \sim N(\underline{m}_t^1, C_t^1)$ where : (for details, see Appendix 6.1A)

$$\underline{m}_t^1 = \underline{a}_t + A_t^1.(\underline{m}_t - \underline{f}_t) \tag{6.17}$$

$$C_t^1 = R_t + A_t^1.(C_t^* - \Sigma_t).A_t^{1^T} \tag{6.18}$$

$$A_t^1 = R_t.F_t.\Sigma_t^{-1} \tag{6.19}$$

$$C_t^* = S_c^{-1}.C_t.(S_c^{-1})^T \tag{6.20}$$

Algorithm Implementation:

The implementation of the above 4-steps algorithm requires the computation of two matrix square roots in step 3, respectively in equations 6.15b and 6.16. These matrix square roots are implemented in two different versions, using respectively two different matrix factorization techniques: the Cholesky decomposition method and the Jacob & related methods. (see Appendix 6.2B for details and references).

The basic difference between these two versions is that , the second one provides not only square root but also the full eigen-structure of V , making possible an additional Principal

Components analysis of the multivariate time series, that is particularly attractive when the dimension d is not small. Another advantage of using the eigen-structure is that this matrix decomposition method is invariant under series permutations , although it takes more processing time.

6.2.4 - Basic Properties :

In order to complete the theory about the scaled multivariate D.L.M. model of this section and also to provide a full insight about its characteristics , we present here four basic results with the corresponding proofs. The first one is a coherence statement , and provides the equivalence between the model formulation of 6.2.1 and the updating equations of section 6.2.2 , in order to unify and validate the whole formulation . The second one , as mentioned before , is the fact that the models of section 5.2 (the common components model and the univariate model) are special cases of the general methods presented here. The third one refers to the important aspect that the associated marginals and conditional distributions also follow the same distributional structure as the joint model. Finally, the last one give us the limiting behaviour of the updating equations and is based on convergence results for dynamic linear models .

i) Specification Equivalence :

The posterior distribution of equations (6.6)-(6.8) given in the model analysis of 6.2.1 , and the posterior distribution of the updating equations (6.13)-(6.15) are equivalent.

Proof:

By equation (6.14a) and using (6.2) & (6.4) , we have:

$A^* = S.\Sigma_t.Q_t^{-1}.S^{-1} = S.A.S^{-1}$, which agrees with (6.6).

Now, by (6.2),(6.4) and (6.5), the expression (6.14) becomes

$C_t = S.\Sigma_t.S - S.\Sigma_t.Q_t^{-1}.\Sigma_t.S = S.[I - A].\Sigma_t.S$, which agrees with (6.6).

Finally, by (6.15b) the coeficient of $\underline{e}_t$, using (6.4) is

$S.(S.Q^{\frac{1}{2}})^{-1} = S.q.S^{-1} = H$, showing that (6.8) and (6.15) are equivalent.

ii) Special Models :

The posterior distribution of the general model of section 6.2 , given by (6.6)-(6.8), under common components assumptions coincides with the posterior distribution of that model. Also, as a consequence , the univariate DLM of section 5.2 is a special case of our general model.

Proof:

Considering the model analysis of 5.2.2 , the adaptive coeficient of (6.5) , will be given by ,

$$A^* = Cov(\underline{\mu}_t, \underline{y}_t).[Var(\underline{y}_t)]^{-1} = (V \otimes F.R_t.F^T).(V \otimes Q_t)^{-1} = k.I$$

where $k = F.R_t.F^T.Q_t^{-1}$, and since $S.A^*.S^{-1} = A^*$, the posterior means of both models coincide. Also, as C_t of (6.6) and (6.14) are the same, and $R_t^* = Var(\underline{\mu}_t/D_{t-1}, V)$ and $Q_t^* = Var(\underline{y}_t/D_{t-1}, V)$ are both proportional to V, then $C_t \propto V$ in both models (with the same proportionality constant). Now, as (6.8) and (6.15) are equivalent , and $Q_t^* = V \otimes Q_t = Q_t.V$ we have, using (6.15b) and (6.16) , that $h_t = \underline{e}_t.Q_t^{-\frac{1}{2}}$; consequently , by (6.15), we get,

$$d_t = \frac{V_t}{n_t} = d_{t-1} + \underline{e}_t.\underline{e}_t^T.Q_t^{-1}$$

which is exactly the updating equation for d_t in the model analysis of 5.2.2 , and the observation updating for both models coincide .

In order to have full equivalence between both algorithms , their time updatings for $\underline{\theta}_t$ should be the same . This means that the equations (6.9)-(6.10) under common components conditions should coincide respectively with equations (5.43)-(5.44) of Appendix 5.2 for the CCM . In fact , by (5.43) we have $M_t^* = G.M_{t-1}$ where these matrices of prior and posterior means are such that $\underline{m}_{t-1}^1 = vec\, M_{t-1}$ & $\underline{a}_t = vec\, M_t^*$. Then ,

$$\underline{a}_t = vecG.M_{t-1} = (I_d \otimes G).vecM_{t-1} = \mathbf{G}.\underline{m}_{t-1}$$

which is exactly eq. (6.9) for $\underline{m}_{t-1} = \underline{m}_{t-1}^1$. Also , by (6.10) , since $\mathbf{R}_t = Var\{\underline{\theta}_t/D_{t-1}, V = I\}$ we have ,

$$\mathbf{R}_t = \mathbf{G}.\mathbf{C}_{t-1}.\mathbf{G}^T + \mathbf{W}_t = (I \otimes G).(I \otimes C_{t-1}).(I \otimes G^T) + (I \otimes W_t)$$
$$= (I \otimes G.C_{t-1}.G^T) + (I \otimes W_t) = I_d \otimes R_t$$

where $R_t = G.C_{t-1}.G^T + W_t$ which is exactly (5.44) .

In practice however we specify $\mathbf{W}_t$ through a discount matrix $\mathbf{B}$ with a discount vector $\underline{b}$ in the diagonal and ones elsewhere . This produces the same results as if we define alternatively $\mathbf{B}^* = (I_d \otimes B)$ where B is a discount matrix for the marginal models . Then ,

$$\mathbf{R}_t = (I_d \otimes G.C_{t-1}.G^T).\mathbf{B} = (I_d \otimes G.C_{t-1}.G^T).(I_d \otimes B) = I_d \otimes R_t$$

where $R_t = GC_{t-1}G^T.B$, which is exactly the practical implementation of (5.44) , and the proof is totally complete .

iii) Approximated Marginal & Conditional Distributions :

Each scalar time series in the d-dimensional vector $\underline{y}_t$ of observations modelled by the scaled multivariate DLM model follows an univariate DLM model. Also, conditional on the values of a given subset of $d - d^*$ time series , the corresponding d^*-dimensional vector $\underline{y}_t^*$ will follow a scaled multivariate DLM model.

Proof : This is a direct consequence of multivariate normal theory.

Corollary: One important consequence of this property is that the model can be used to define not only joint and marginal forecasting functions but also Conditional Forecasting Functions , which is an extremely attractive feature of such multivariate models.

iv) Convergence result :

If a given scaled multivariate DLM is constant and observable , (that is , the matrix T given by (6.21) has full rank) then , using arguments analogous to those in Quintana(1987) , the following limiting relationships about adaptive coeficients and variance-covariance matrices are conjectured ,

$$\lim_{t \to \infty} \{A_t^*, C_t^1, R_t, R_t^*, Q_t^*\} = \{A^*, C^1, R, R^*, Q^*\}$$

with the following relations :

$$R = G.C^1.G^T + W \quad ; \quad A^* = R^*.Q^{*^{-1}} \quad ; \quad Q^* = R^* + V$$

Comment :

For convergence results about observable constant DLM's , also valid for multivariate constant DLM's , see Harrison(1985) and West & Harrison(1989) .

In the particular case of common components , the limiting results above coincide with the limiting equations for the CCM , as given in Quintana(1987).

6.3 - An Alternative Model Representation .

Although the method proposed previously in the last section is perfectly valid for modelling and forecasting purposes , the model is basically defined in the $\underline{\mu}$ parametrization where the variance structure is specified , and full unconditional distribution for $\underline{\theta}$ is not available .

In fact , given an initial probabilistic information for $(\underline{\theta}_0/D_0\,, V = S_0^2)$ where S_0 is a reference variance matrix (see step 3 of section 6.2.3) , all we can get from the model with respect to $\underline{\theta}_t$ is the distribution of $(\underline{\theta}_t/D_t\,, V = S_R^2)$, restricted on a reference variance matrix $S_R = S_0$.

Consequently , the possibility of intervention in the $\underline{\theta}$ prior for instance is more restricted and an alternative procedure where full (unrestricted) prior and posterior distributions are available is desired . That is , we want the posterior distribution $(\underline{\theta}_t/D_t\,, V)$ for a general and unrestricted variance matrix V .

In this way , we will keep the basic model formulation for $\underline{\mu}_t$ as in section 6.2 but an alternative procedure to bring back the information from $\underline{\mu}_t$ to $\underline{\theta}_t$ will be introduced , instead of using Conditional Independence relations , that will make a direct new model formulation for $\underline{\theta}_t$ possible .

In order to make easier the full understanding of the new model formulation and analysis introduced in section 6.3.3 , we present two introductory sub-sections - 6.3.1 and 6.3.2 - with basic definitions and concepts as well as some theoretical background .

6.3.1 - Basic Definitions and Concepts .

We consider here an alternative way of structuring the prior covariance matrix R under the same basic model of last section . But now , dimension compatibility is reached not contracting R to the S dimension , but expanding S to a R compatible dimension through the observability matrix T defined by ,

$$T = \begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_k \end{pmatrix} = \begin{pmatrix} F \\ FG \\ \vdots \\ FG^{k-1} \end{pmatrix} \tag{6.21}$$

with dimension kdxp , where k is the maximum parametric dimension of the marginal models , d is the dimension of the observations and p is the process parameter dimension .

In order to understand the meaning and usefulness of this matrix T , lets see how it works in practice as a transformation matrix from the space generated by $\underline{\theta}_t$ to the space generated by $\underline{\mu}_t$ and vice-verse through a very simple illustrative example .

Example : Without loss of generality with respect to the aspects we want to focus attention , consider the unidimensional linear growth model with zero noise (otherwise , consider expected values of parameters and observations) defined by ,

$$y_t = \mu_t \tag{6.22a}$$

$$\begin{pmatrix} \mu_t \\ \beta_t \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} . \begin{pmatrix} \mu_{t-1} \\ \beta_{t-1} \end{pmatrix} \tag{6.22b}$$

where μ_t (mean response parameter) represents the process level and β_t is the slope parameter. We notice that the information about $\underline{\theta}_t$ is transferred (partially) to μ_t through the relation ,

$$\mu_t = F.\underline{\theta}_t = (1 \quad 0) . \begin{pmatrix} \mu_t \\ \beta_t \end{pmatrix} \tag{6.23}$$

but the reciprocal (the 'bring back' of information from μ_t to $\underline{\theta}_t$) is not obvious since F is not invertible . In order to overcome this difficulty we rewrite (6.23) with t substituted by t+1 ,

$$\mu_{t+1} = F.\underline{\theta}_{t+1} = FG.\underline{\theta}_t = (1 \quad 0) . \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} .\underline{\theta}_t \tag{6.23a}$$

and add (6.23a) to (6.23) , giving

$$\begin{pmatrix} \mu_t \\ \mu_{t+1} \end{pmatrix} = \begin{pmatrix} F \\ FG \end{pmatrix} .\underline{\theta}_t = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} .\underline{\theta}_t \tag{6.24}$$

Now , the inverse transformation is possible , giving

$$\underline{\theta}_t = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}^{-1} . \begin{pmatrix} \mu_t \\ \mu_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} . \begin{pmatrix} \mu_t \\ \mu_{t+1} \end{pmatrix} = \begin{pmatrix} \mu_t \\ \beta_t \end{pmatrix} \tag{6.25}$$

Basically , what happens is that the information about the slope β_t lost after the contraction transformation F can be recovered considering the additional transformation FG . In a more general situation where we have more parameters and $\underline{y}_t$ is a vector , the same ideas apply and the transformation from $\underline{\theta}_t$ to $\underline{\mu}_t$ is the one defined by equation 6.21 .

Definitions : Now , the expanded scale matrix S_T of dimension kdxp is defined by ,

$$S_T = \begin{pmatrix} S^*.T_1 \\ S^*.T_2 \\ \vdots \\ S^*.T_k \end{pmatrix} \tag{6.26}$$

where S^* is such that $V = S^*.V_{t-1}.S^*$, and the scaled version of R will be defined by ,

$$R_T = T^-.S_T.R.S_T'.T^{-'} \tag{6.27}$$

where $R = Var[\theta_t/D_{t-1}, V = V_{t-1}]$ and T^- is the Penrose generalized inverse of the observability matrix T , which will coincide with the standard inverse for common components models .

6.3.2 - Theoretical background

Although it may seem arbitrary ,in fact, the variance structure (6.26)-(6.27) was designed in order to hold certain specific relations as we can see through the following analysis :

i) Rewriting the equation (6.27) as $T.R_T.T' = S_T.R.S_T'$ and using the definitions (6.21) and (6.26) , we have

$$\begin{pmatrix} F' \\ F'.G \\ \vdots \end{pmatrix}.R_T.(F' \quad F'G \quad ..) = \begin{pmatrix} S^*.F' \\ S^*.F'G \\ \vdots \end{pmatrix}.R.(S^*.F' \quad S^*.F'G \quad ..) \tag{6.28}$$

which immediately implies that

$$Var[F^T.G^j.\underline{\theta}_t/D_{t-1}, V = S^*.V_{t-1}.S^*] = S^*.F^T.G^j.R_t.G'^j.F.S^* \tag{6.29}$$

j=0,1,..,k-1 , where , for j=0 , it gives

$$Var[\underline{\mu}_t/D_{t-1}, V = S^2] = S^*.F^T.R_t.F.S^* = S^*.\Sigma_t^*.S^* = S.\Sigma_t.S \tag{6.30}$$

since $\Sigma_t^* = Var[\underline{\mu}_t/D_{t-1}, V = V_{t-1}] = V_{t-1}^{\frac{1}{2}}.\Sigma_t.V_{t-1}^{\frac{1}{2}}$ and $S = S^*.V_{t-1}^{\frac{1}{2}}$ by construction.

We notice that (6.30) is exactly the same variance structure considered in the last section, as we can see from equations (6.12)-(6.12a) . As a consequence (see property ii of section 6.2.3) , in the case of $j = 0$, the variance structure (6.29) is in accordance with the conjugate analysis for the common components model. That is the case , for instance , of the multivariate steady model - a multidimensional generalization of the univariate steady model presented by Harrison & Stevens (1976) - where each marginal univariate series is described by only one parameter (k=0) , the level process , and then $j = 0$.

ii) Now , to see that not only for $j = 0$ but for any $j = 0, 1, .., k-1$, the relations (6.29) are in exact accordance with the conjugate analysis of the common components model of section 5.2.2 , we proceed as follows :

From the left hand side of (6.29) , we have ,

$$Var[\mathbf{F}^T.\mathbf{G}^j.\underline{\theta}_t/D_{t-1}, V = S^*.V_{t-1}.S^*] = \mathbf{F}'.\mathbf{G}^j.\mathbf{R}_t.\mathbf{G}'^j.\mathbf{F} \tag{6.31}$$

Under common components conditions $(\mathbf{F} = I \otimes F \;;\; \mathbf{G} = I \otimes G \;;\; \mathbf{R}_t = V \otimes R_t)$ and using operating properties of Kronecker products such as $(A \otimes B).(C \otimes D) = A.C \otimes B.D$; $(A \otimes B)^k = A^k \otimes B^k$; $(A \otimes B)' = A' \otimes B'$, etc , the equation (6.31) became

$$Var[\mathbf{F}'.\mathbf{G}^j.\underline{\theta}_t/.] = V \otimes F'G^j.R_t.G'^jF = c.V = c.S^*.V_{t-1}.S^* \tag{6.32}$$

where $c = F'.G^j.R.G'^j.F$ is a scalar.

Now , from the right hand side of (6.29) , and under the same conditions (common components) , we get ,

$$Var[\mathbf{F}'.\mathbf{G}^j.\underline{\theta}_t/.] = S^*.V_{t-1} \otimes F'G^j.R_t.G'^jF.S^* = c.S^*.V_{t-1}.S^* \tag{6.33}$$

Finally , comparing (6.32) with (6.33) we see that these variances coincide exactly .

6.3.3 - Model Formulation and Analysis

Definition : An alternative formulation of the multivariate DLM model of section 6.2 for a vector of observations $\underline{y}_t$ of dimension d made at intervals , at times t=1,2,.. is defined by the equations (5.1)-(5.2) with the following distributional structure :

i) prior distribution for V : $(V^{-1}/D_{t-1}) \sim W(d_{t-1}; n_{t-1})$ where d_{t-1} and n_{t-1} are respectively the shape parameter and the d.f. of the Wishart distribution such that $V_{t-1} = d_{t-1}/n_{t-1}$.

ii) joint prior distribution for $\underline{\theta}_t$ and $\underline{y}_t$:

$$\begin{pmatrix} \underline{\theta}_t \\ \underline{y}_t \end{pmatrix} \Big| \; D_{t-1}, V = S^*.V_{t-1}.S^* \sim N\left\{ \begin{pmatrix} \underline{a}_t \\ \underline{f}_t \end{pmatrix}; \begin{pmatrix} R_T & R_T.F \\ . & F^T.R_T.F + V_{t-1} \end{pmatrix} \right\} \tag{6.34}$$

where $R_T = T^-.S_T.R.S_T'.T'^-$, with T and S_T as defined in (6.21) & (6.26) respectively and $R = Var[\underline{\theta}_t/D_{t-1}, V = V_{t-1}]$

iii) As a direct consequence of the definition above given by (i) & (ii) , we have also that the joint prior distribution for $\underline{\mu}_t$ and $\underline{y}_t$ is given by :

$$\begin{pmatrix} \underline{\mu}_T \\ \underline{y}_t \end{pmatrix} \Big| \; D_{t-1}, V = S.S' = S^*.V_{t-1}.S^* \sim N\left\{ \begin{pmatrix} F^T.\underline{a}_t \\ \underline{f}_t \end{pmatrix}; \begin{pmatrix} S.\Sigma_t.S' & S.\Sigma_t.S' \\ . & S.Q_t.S' \end{pmatrix} \right\} \tag{6.35}$$

where all elements in the covariance matrix of (6.35) can be equivalently written with the star superscript since , for instance ,

$$Q_t^* = q.q' = Var[\underline{y}_t/D_{t-1}, V = V_{t-1}] = V_{t-1}^{\frac{1}{2}}.Q_t.V_{t-1}^{\frac{1}{2}}$$

and then ,

$$S^*.Q_t^*.S^* = S.V_{t-1}^{-\frac{1}{2}}.V_{t-1}^{\frac{1}{2}}.Q_t.V_{t-1}^{\frac{1}{2}}.V_{t-1}^{-\frac{1}{2}}.S = S.Q_t.S$$

Model Analysis :

i) Posterior distribution for V :

From the joint prior distribution given in (iii) above , we have ,

$$(\underline{e}_t/D_{t-1}, V = S^2 = S^*.V_{t-1}.S^*) \sim N[\underline{0}; S.Q_t.S] \tag{6.36}$$

Or, equivalently , $(H.\underline{e}_t/D_{t-1}, V = S^2) \sim N(\underline{0}; V)$ where $H = S.(S^*.q)^{-1} = S.q^{-1}S_{t-1}.S^{-1}$. Now , supposing that H is locally constant around $V = V_{t-1} = S_{t-1}.S_{t-1}^{'}$, and , as a consequence , approximately equal to $S_{t-1}.q^{-1}$, we have , using standard conjugate analysis , the following posterior for V : $(V^{-1}/D_t) \sim W(d_t; n_t)$ where ,

$$d_t = d_{t-1} + h_t.h_t^{'} \tag{6.37a}$$

$$h_t = S_{t-1}.q^{-1}.\underline{e}_t \tag{6.37b}$$

$$n_t = n_{t-1} + 1 \tag{6.37c}$$

which coincides with the posterior distribution for V given by the equations (6.15)-(6.16) .

ii) Posterior distribution for $\underline{\mu}_t$:

From the joint prior distribution for $\underline{y}_t$ and $\underline{\mu}_t$, we have that the adaptive coeficients matrix A^* is given by ,

$$\begin{aligned} A^* &= Cov(\underline{y}_t, \underline{\mu}_t/D_{t-1}, V).Var(\underline{y}_t/D_{t-1}, V) \\ &= S.\Sigma_t.S^{'}.(S.Q.S^{'})^{-1} = S.\Sigma_t.Q_t^{-1}.S^{-1} = S.A_t.S^{-1} \end{aligned} \tag{6.38}$$

As a consequence , the posterior distribution for $\underline{\mu}_t$ will be given by

$$(\underline{\mu}_t/D_t, V) \sim N[F^{'}.\underline{a}_t + A^*.\underline{e}_t \quad ; \quad S^*.(I - A^*).\Sigma_t^*.S^*\,] \tag{6.39}$$

which coincides with the posterior distribution of the model formulation of section 6.2 .

(iii) Posterior distribution for $\underline{\theta}_t$: From the joint prior distribution for $\underline{y}_t$ and $\underline{\theta}_t$, the adaptive coeficient matrix A_T is given by ,

$$A_T = Cov(\underline{y}_t, \underline{\theta}_t/D_{t-1}, V).Var(\underline{y}_t/D_{t-1}, V) = R_T.F.Q_t^{-1} \tag{6.40}$$

where $Q_T = F^{'}.R_T.F + V_{t-1}$. As a consequence , conditional on V , the posterior distribution for $\underline{\theta}_t$ will be given by $(\underline{\theta}_t/D_t, V) \sim N(\underline{m}_t; C_t)$ where :

$$\underline{m}_t = \underline{a}_t + A_T.\underline{e}_t \tag{6.41a}$$

$$C_t = (I - F.A_T).R_T \tag{6.41b}$$

Also , the unconditional posterior distribution for $\underline{\theta}_t$ will be given by ,

$$p(\underline{\theta}_t/D_t) = \int p(\underline{\theta}_t/D_t, V).p(V/D_t)\,dV$$

where the densities for $\underline{\theta}_t$ and V in the integrand were defined before . Solving this integral in V , we get a multivariate t distribution for the unconditional posterior distribution of $\underline{\theta}_t$, with parameters $\underline{m}_t$, C_t and n_t , as defined before.

6.3.4 - Updating Equations :

As a consequence of the definitions and analysis of sections 6.3.1 to 6.3.3 , we have the following Updating Equations defined directly in terms of the $\underline{\theta}_t$ parameter :

notation : $(\underline{\theta}_{t-1}/D_{t-1}, V) \sim N(\underline{m}_{t-1}, C_{t-1})$

i) Time Updating : $(\underline{\theta}_t/D_{t-1}, V) \sim N(\underline{a}_t, R_t)$ where :

$$\underline{a}_t = G.\underline{m}_{t-1} \tag{6.42a}$$

$$R_t = G.C_{t-1}.G^T + W_t \tag{6.42b}$$

In practice , W_t is specified through a discounting factor $0 \leq \delta \leq 1$ - see Harrison & West(1987)

.

ii) Observation Updating for V : $(V^{-1}/D_t) \sim W(d_t; n_t)$ where ,

$$d_t = d_{t-1} + h_t.h_t^T \tag{6.43}$$

$$h_t = V_{t-1}^{\frac{1}{2}}.[Q_t^{\frac{1}{2}}]^{-1}.\underline{e}_t \tag{6.43a}$$

$$Q_t = F^T.R_t.F + V_{t-1} \tag{6.43b}$$

$$\underline{e}_t = \underline{y}_t - F^T.\underline{a}_t \tag{6.43c}$$

$$n_t = n_{t-1} + 1 \tag{6.44}$$

$$V_t = n_t^{-1}.d_t \tag{6.45}$$

$$S_t = V_t^{\frac{1}{2}} \tag{6.45a}$$

iii) Scaling and Observation Updating for $\underline{\theta}_t$: $(\underline{\theta}_t/D_t, V) \sim N(\underline{m}_t, C_t)$ where :

$$\underline{m}_t = \underline{a}_t + A_T.\underline{e}_t \tag{6.46}$$

$$A_T = R_T.F.Q_T^{-1} \tag{6.46a}$$

$$C_t = (I - F.A_T).R_T \tag{6.47}$$

$$R_T = T^{-}.(S_T.R_t.S_T).T^{-'} \tag{6.47a}$$

$$Q_T = F^T.R_T.F + V_t \tag{6.48}$$

$$S^* = S_t.S_{t-1} \tag{6.49}$$

$$S_T = \begin{pmatrix} S^*.T_1 \\ : \\ S^*.T_k \end{pmatrix} \tag{6.50}$$

$$T_j = F'.G^{j-1} \qquad ,j = 1,.,k \tag{6.50a}$$

Important Implementation aspects :

i) This algorithm was implemented considering two different numerical methods for matrix square root evaluation (equations 6.25 and 6.25a) : the Cholesky decomposition and the Jacobi method . Also , numerical methods for generalized inverse matrix evaluation are used in the implementation of the equation 6.47a . For details and references about these numerical procedures , see Appendices 6.2a and 6.2b .

ii) Although there is theoretical equivalence between the model structure of this section and the former formulation of section 6.2 (and consequently the conjugate analysis of the CCM is a particular case) , its practical implementation using discount factors procedures in the time updating stage does not necessarily keep this equivalence . In fact , given a discount matrix **B** , under common components conditions , we have

$$\begin{aligned} \mathbf{R}_t &= \mathbf{G}.\mathbf{C}_{t-1}.\mathbf{G}^T.\mathbf{B} = (I \otimes G_t).(V_{t-1} \otimes C_{t-1}).(I \otimes G^T).\mathbf{B} \\ &= (V_{t-1} \otimes G.C_{t-1}.G^T).\mathbf{B} \end{aligned}$$

The usual definition of **B** with a discount vector $\underline{b}$ in the diagonal and ones elsewhere (diagonal discount) clearly destroy the mentioned equivalence , although this is still the recommended practical procedure .

Such equivalence however can be obtained using the following alternative specification for the discount matrix . Consider a common discount factor $0 < \beta \leq 1$ for all matrix elements

(variances or covariances) such that $\mathbf{B} = \frac{1}{\beta}\begin{pmatrix} 1 & .. & 1 \\ : & :: & : \\ 1 & .. & 1 \end{pmatrix}$. As a consequence , we have , $\mathbf{R}_t = V_{t-1} \otimes R_t$, where $R_t = G.C_{t-1}.G^T/\beta$, which is the standard practical variance time updating for the CCM .

For this reason , we have implemented the updating equations of this section using both discount options : block diagonal discount and common discount .

6.4 - An Application with real data

We present here one example of application with real data where the algorithm of section 6.3 for updating of the general multivariate DLM is used .

6.4.1 - Data Modelling and Analysis :

The data in question is a three-dimensional time series of soft wheat prices per 100Kg in the three major European economies : Great Britain , France and West Germany . The series consist of monthly average prices , expressed in European Currency Units - ECU , from January of 1982 to April of 1988 . The reference source of this data is EUROSTAT , and consists originally of agricultural price series from the CRONOS DATABANK .

Commodity price series in general present annual seasonality and , in the case of the present data , this can be observed in fig. 6.1 . In fact , a more carefull analysis of the seasonal pattern present in the data using a Periodogram (see figs 6.4 , 6.5) , suggests the presence of the first two harmonics (12 months and 6 months seasonality) in both the DTPRICE and FRPRICE series . The GBPRICE series however does not present any significant harmonic at all , and consequently , no seasonal component is considered for this series .

These harmonic analysis results were obtained through the package B.A.T.S. (Bayesian Analysis of Time Series) and the Periodogram is essentially the squared magnitude of the discrete Fourier Transform of the data (each individual scalar time series) $D_t = \{y_1, y_2, \cdots, y_t\}$, i.e.

$$C(w) = \frac{1}{t}.\left|\sum_{j=1}^{t} y_j.e^{iwt_j}\right|^2 = \frac{1}{t}.[\,R^2(w) + I^2(w)\,] \qquad \text{with}$$

$$R(w) = \sum_{j=1}^{t} y_j.\cos wt_j \quad \& \quad I(w) = \sum_{j=1}^{t} y_j.\sin wt_j$$

where $w = \frac{2\pi}{T}$ is called respectively first harmonic , second harmonic , ... etc for T = 12 , 6 ,... (it is considered monthly data) , and $C(w)$ will give the proportion of variation in the data explained by each individual harmonic .

A model is specified (multivariate DLM) with a linear trend component and a 2-harmonic structure for each series , except the GBPRICE series where there is only the trend term .

At this point , it is interesting to mention two aspects related with model selection , or more specifically , with the identification of the seasonal structure that should be present in the model . The first one is that if we want a more parsimonious model , we could perhaps drop the second harmonic in the DTPRICE series since it has a relatively small significance as showed in fig 6.4 . The second aspect is related with the data measurement unit considered , i.e. , E.C.U. The model we have chosen , and particularly the seasonal structure , would be different if the prices were given in their national currencies - Pounds , Francs and Marks . In this case , as suggested similarly by a spectral analysis of each new time series , only the FRPRICE (in Francs) series would follow the same previous two-harmonic structure ; the DTPRICE (in Marks) series would need only the first harmonic and the GBPRICE (in Pounds) series would need one or perhaps two harmonics in its seasonal structure .

It is clear that , because the seasonal structure is not common to all series , the model does not have more an overall common component structure , but a general multidimensional DLM form . Concretely , the model under consideration can be written as

$$\underline{y}_t = \mathbf{F}.\underline{\theta}_t + \underline{v}_t \quad , \quad \underline{v}_t \sim N(\underline{0}, V)$$
$$\underline{\theta}_t = \mathbf{G}.\underline{\theta}_{t-1} + \underline{w}_t \quad , \quad \underline{w}_t \sim N(\underline{0}, W)$$

with the design elements $\mathbf{F}$ & $\mathbf{G}$ specified as $\mathbf{F} = diag\{F_1, F_2, F_3\}$ where

$$F_1 = F_2 = (1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0) \quad \& \quad F_3 = (1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$$

and the system matrix is $\mathbf{G} = diag\{G_1, G_2, G_3\}$ where

$$G_1 = G_2 = diag\left\{\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}; \begin{pmatrix} .866 & 0.5 \\ -0.5 & .866 \end{pmatrix}; \begin{pmatrix} 0.5 & .866 \\ -.866 & 0.5 \end{pmatrix}\right\} \quad \& \quad G_3 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

In order to make use of the algorithm of last section , relatively non-informative initial priors (standard priors with large values for the parameters associated with uncertainty) are set where $C_0 = 100.I$ and $n_0 = 0.001$. The one-step-ahead forecasting residuals from this model

is presented in fig 6.2 , which suggests , with the exception of a few points , a good agreement between model and data . The correlation between the component series are presented in fig 6.3 , which show a high correlation between the series of German and French prices (around 0.8 or 0.9) and a weak correlation between these series and the GBPRICE series . In fact , the original series suggest this sort of behavior (without quantifications of course) , which reflects a more independent or peculiar position of Great Britain in terms of economic policy , particularly in the case of agricultural products .

6.4.2 - A Conditional Analysis :

Another interesting aspect we can explore in our analysis is related to the way the data is sequentially available in time . In general , we can notice that the agricultural prices for West Germany are available in EUROSTAT at least 1 month ahead in relation to the prices in the other two countries . For instance , if today the more recent data available in EUROSTAT about commodity prices (wheat in particular) in England or France is relative to March/89 , the same data for West Germany is available till April or May/89 for instance .

In general , what happens with our data , say $\underline{y}_{t-1} = \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \end{pmatrix}$ is that , when we are using data till time t-1 to make 1-step-ahead forecast about $\underline{y}_t$, in fact , y_{1t} is available (or another subset of $\underline{y}_t$) and we can use this information to improve our view about the future values of the other components of $\underline{y}_t$. Then , using the notation of the algorithm of section 6.3 , the predictive distribution for $\underline{y}_t$ is given by

$$(\underline{y}_t / D_{t-1}, V) \sim N(\underline{f}_t = F^T . \underline{a}_t \quad , \quad Q_t = F^T . R_t . F + V_{t-1})$$

Now , given that y_{1t} is known at time t-1 , the conditional predictive distribution for $\underline{y}_t^* = \begin{pmatrix} y_{2t} \\ y_{3t} \end{pmatrix}$ is given by $(\underline{y}_t^* / D_{t-1}, y_{1t}, V) \sim N(\underline{f}_t^*, Q_t^*)$ where :

$$\underline{f}_t^* = \underline{f}_{2t} + Q_{21} . q_{11}^{-1} . (y_{1t} - f_{1t})$$

$$Q_t^* = Q_{22} - Q_{21} . q_{11}^{-1} . Q_{12}$$

Notation : $\underline{f}_t = \begin{pmatrix} f_{1t} \\ \underline{f}_{2t} \end{pmatrix}$ & $Q_t = \begin{pmatrix} q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}$

. As a consequence , conditional on a given value of y_{1t} , we can now define one-step-ahead forecasting residuals for the other two series as $\underline{e}_t^* = \underline{y}_t^* - \underline{f}_{2t}$. These conditional residuals are shown in fig 6.7 , which can be compared with the usual residuals .

Appendix 6.1a - Some Conditional Independence Results

Consider the following basic definitions involving the random vectors X , Z & U and the concept of Conditional Independence - C.I. for short :

(i) X is Conditionally Independent of U , given Z , if and only if , X given Z has the same probability distribution as X given Z and U .

notation : $X \perp U/Z \leftrightarrow X/Z \equiv X/Z,U$

(ii) X and U are Conditionally Independent given Z if and only if, given Z , X is Conditionally Independent of U and U is Conditionally Independent of X.

Now, lets consider some basic results involving normality and conditional independence:

LEMMA: If the normal random vectors X , U & Z are such that X and U are Conditionally Independents given Z , then , we have

$$Cov[X,U] = Cov[X,Z].[Var(Z)]^{-1}.Cov[Z,U]$$

Proof : Considering the notation

$$Var\begin{pmatrix} X \\ U \\ -- \\ Z \end{pmatrix} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XU} & | & \Sigma_{XZ} \\ -- & \Sigma_{UU} & | & \Sigma_{UZ} \\ --- & --- & -- & --- \\ -- & -- & | & \Sigma_{ZZ} \end{pmatrix}$$

from standard multivariate normal theory , we have

$$Var\begin{pmatrix} X & | & Z \\ U & | & . \end{pmatrix} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XU} \\ -- & \Sigma_{UU} \end{pmatrix} - \begin{pmatrix} \Sigma_{XZ} \\ \Sigma_{UZ} \end{pmatrix}.\Sigma_{ZZ}^{-1}.(\Sigma_{ZX} \quad \Sigma_{ZU})$$

Then , since Cov (X ,U / Z) = 0 by hypothesis , we have

$$\Sigma_{XU} - \Sigma_{XZ}.\Sigma_{ZZ}^{-1}.\Sigma_{ZU} = 0 \qquad \text{, or equivalently}$$

$$Cov[X,U] = Cov[X,Z].[Var(Z)]^{-1}.Cov[Z,U]$$

THEOREM: If X , Z and U have a joint normal distribution and, given Z, X and U are Conditionaly Independent, then

$$\begin{pmatrix} X & | & U = u \\ Z & | & . \end{pmatrix} \sim N\left\{\begin{pmatrix} m_X^* \\ m_Z^* \end{pmatrix}; \begin{pmatrix} C_X^* & A^*.C_Z^* \\ . & C_Z^* \end{pmatrix}\right\}$$

where:

$$A^* = Cov(X,Z).[Var(Z)]^{-1}$$

$$m_X^* = m_X + A^*.(m_Z^* - m_Z) \quad (1)$$

$$C_X^* = C_X + A^*.(C_Z^* - C_Z).A^{*T} \quad (2)$$

Proof:

(i) Let $A = Cov(Z,U).[Var(U)]^{-1}$. Then, using Lemma 1 ,

$$A^*.A = Cov(X,Z).[Var(Z)]^{-1}.Cov(Z,U).[Var(U)]^{-1} = Cov(X,U).[Var(U)]^{-1}$$

Then, as $m_Z^* = m_Z + A.(u - m_U)$ we have:

$$m_X^* = m_X + A^*.A.(u - m_U) = m_X + A^*.(m_Z^* - m_Z) \quad (1)$$

(ii) $C_Z^* = C_Z - A.C_U.A^T$. Then, as $C_X^* = C_X - A^*.A.C_U.A^T.A^{*T}$, we have

$$C_X^* = C_X + A^*.(C_Z^* - C_Z).A^{*T} \quad (2)$$

Finally, lets consider the specific case of multivariate D.L.M's , where $X = \underline{\theta}$, $Z = \underline{\mu} = F.\underline{\theta}$ and $U = \underline{y}$. At each time t , given $\underline{\mu}_t = F.\underline{\theta}_t$, $\underline{\theta}_t$ and $\underline{y}_t$ are conditionaly independent by definition . Using our usual notation , the prior moments are $(\underline{a}_\theta, R_\theta)$ and $(\underline{a}_\mu, R_\mu)$ and the conditionaly independent posterior moments are $(\underline{m}_\theta, C_\theta)$ and $(\underline{m}_\mu, C_\mu)$. Then, the relations (1) and (2) can be rewritten as:

$$\underline{m}_\theta = \underline{a}_\theta + A^*.(\underline{m}_\mu - \underline{a}_\mu) \quad (1')$$

$$C_\theta = R_\theta + A^*.(C_\mu - R_\mu).A^{*T} \quad (2')$$

where $A^* = Cov(\theta,\mu).[Var(\mu)]^{-1} = F^T.R_\theta.R_\mu^{-1}$

Appendix 6.1b - Some Sensitivity Analysis Results

notation : A is a dxd real matrix and S is such that $S.S^T = V$ where V is a dxd covariance matrix.

Lets study the sensitivity of $T = S.A.S^{-1}$ to variations or perturbations in the matrix S or V. Initially we will consider the special cases of first order (linear or proportional) and second

order perturbations in S using ordinary tools, and after , the more general case, using tensor differential calculus.

a) special perturbations : 1^{st} & 2^{nd} order cases.

One very important special case is when all elements of V vary or are perturbated proportionally to the same constant λ^2 say. Or, equivalently, when all elements of S vary proportionally to the same constant λ. In this case, as

$$(\lambda S).A.(\lambda S)^{-1} = S.A.S^{-1}$$

there is no resulting perturbation in T.

Now, in the case of 2^{nd} order perturbations, given by

$$S \longrightarrow S + \epsilon.S_1 + \frac{1}{2}.\epsilon^2.S_2 + O(\epsilon^3)$$

it will induce a perturbation in T , given by

$$T \longrightarrow T + \epsilon.T_1 + \frac{1}{2}.\epsilon^2.T_2 + O(\epsilon^3)$$

where:

$$T_1 = R_1.T - T.R_1 \quad ; \quad T_2 = (R_2.T - T.R_2) - 2.T_1.R_1$$

$$R_1 = S_1.S^{-1} \quad ; \quad R_2 = S_2.S^{-1}$$

obviously, when $S = S_1$ and $S_2 = 0$ we have the proportional case.

b) sensitivity analysis - general case :

Considering the matrix differential of $T = S.A.S^{-1}$, we have:

$$dT = (dS).A.S^{-1} + S.A.d(S^{-1}) \text{ , or}$$

$$d(vecT) = [(A.S^{-1})^T \otimes I_d].(vecdS) + [I_d \otimes S.A].(vecd(S^{-1}))$$

or, equivalently ,

$$d(vecT) = \{[(A.S^{-1})^T \otimes I_d] - [(S^{-1})^T \otimes S.A.S^{-1}]\}d(vecS)$$

where the dxd matrix between curly brackets is the Derivative Matrix. This matrix , express all the information about how sensitive is $S.A.S^{-1}$ to variations in S. However, this information can be summarized in a simple way through its Euclidean Norm, as follows :

$$|D| = ||[(A.S^{-1})^T \otimes I_d] - [(S^{-1})^T \otimes T]||$$

or,using well known inequalities, we have

$$|D| \leqslant |(A.S^{-1})^T \otimes I_d| + |(S^{-1})^T \otimes T|$$

and, after some algebra , we get

$$|\frac{\partial T}{\partial S}| \leqslant d.tr(A.V^{-1}.A^T) + tr(V^{-1}).tr(V.A.V^{-1}.A^T).$$

For more details about matrix or tensor differential calculus, a good reference is Magnus & Neudecker(1988).

Appendix 6.2a - Generalized Inverse Matrix Techniques

I - Basic Definitions and Properties :

DEF.1 : A matrix is Hermitian if it equals its own Hermitian transpose (or complex conjugate transpose $A^H = \overline{A}^T$) ; that is , A is Hermitian if $A = A^H$. In practice , a Hermitian matrix is the counterpart of a symmetric matrix when A has complex elements . When A is real , both concepts coincide . One important property of such matrices is that the eigenvalues of a Hermitian matrix are real .

DEF.2 : The (Moore-Penrose) generalized inverse (or pseudo-inverse) of a matrix A , not necessarily square , is a matrix A^- that satisfies the following conditions :

$$i)\ A.A^- \quad \text{and} \quad A^-.A \quad \text{are Hermitian.}$$

$$ii)\ A.A^-.A = A$$

$$iii)\ A^-.A.A^- = A^-.$$

In the case where only the condition ii) is satisfied , A^- is called a conditional inverse of A ; some authors , such as Rao (1973) , define g-inverses in this more general way (for specific applications where uniqueness is not necessary) .

PROP.1 (Existence & Uniqueness) : For every matrix A of dimension nxm , the M-P generalized inverse A^- exists and is unique (with dimension mxn) .

PROP.2 (Operational Properties) :

$$i)\ A^- = A^{-1} \quad \text{for a non-singular square matrix A}$$

$$ii)\ (A^-)^- = A \quad ; \quad (A^H)^- = (A^-)^H$$

$$iii)\ (kA)^- = \frac{1}{k}.A^- \quad \text{for} \quad k \neq 0 \quad ; \quad 0^- = 0 \quad (\text{null matrix})$$

iv) When the columns of A form a linearly independent set of vectors , we have , $A^- = (A^H.A)^{-1}.A^H$.

II - Computing MP-generalized Inverses :

Although the last operational property (iv) provides a simple and practical way of calculating generalized inverses , it is restricted to the case of full rank matrices . In our context of multivariate DLM's of this chapter , A is an observability matrix T (see definition 6.21) where for most practical situations we have full rank . However , there are models where T doesn't have full rank and , a more general matrix inversion procedure can be necessary .

An iterative scheme to find the MP-inverse of a matrix is given by the following Theorem [Graybill (1983)] :

Theorem : Let A be an mxn matrix , let A_{t-1} be the mx(t-1) matrix consisting of the first t-1 columns of A and let a_t be the t^{th} column of A . Then , we have

$$A^- = \begin{pmatrix} A_{t-1}^- - A_{t-1}^-.a_t.b_t^- \\ b_t^- \end{pmatrix} \qquad \text{where,}$$

$b_t = (I - A_{t-1}.A_{t-1}^-).a_t \qquad$ if $\qquad a_t \neq A_{t-1}.A_{t-1}^-.a_t \qquad$ (case I)

$b_t = \dfrac{[1 + a_t^T.(A_{t-1}.A_{t-1}^T)^-.a_t](A_{t-1}.A_{t-1}^T)^-.a_t}{a_t^T.(A_{t-1}.A_{t-1}^T)^-.(A_{t-1}.A_{t-1}^T)^-.a_t}$ if $a_t = A_{t-1}.A_{t-1}^-.a_t \;\&\; a_t \neq 0$ (II)

$b_t = 0 \qquad$ if $\qquad a_t = 0 \qquad$ (case III) .

Appendix 6.2b - Factorization Methods for p.d. Matrices

In the methodology introduced in this chapter , two basic factorization algorithms and related techniques were used: the Cholesky decomposition method, that gives the square roots of a symmetric matrix , and the Jacob method, that provide full eigen-structure , and consequently square roots also. We present here the basic definitions and properties about these and related methods. A more detailed description of these techniques and full references can be found for instance , in Bierman,G.J.(1977) or Press,W.H. at all (1986)

A) Matrix Square Roots & the Cholesky Decomposition Algorithm

Definition: If the positive definite matrix V can be written as $V = S.S^T$, with S a square matrix, we say that S is a square root of V. This definition, although frequently used, is not universal. Some authors allow S to be rectangular , while others restrict S to be symmetric.

Basic Properties:

i) Unicity: Square Roots matrices, when they exist, are non-unique. If S is one square root of V , and P is orthogonal ($P.P^T = I$) , then S.P is also a square root of V.

ii) Existence: Every P.D. matrix has a square root. This can be verified by the elementar process of completing squares in the corresponding quadratic forms.

iii) Triangular Factorization: If S1 and S2 are two lower triangular factorizations of V , then $S1 = S2.diag(\pm 1, ..., \pm 1)$, i.e. $S(j,j) = \pm[V(j,j)]^{\frac{1}{2}}$. From now on, we will consider the unique square root corresponding to positive diagonal elements.

Lower Triangular Cholesky Decomposition

If V is P.D., it has a lower triangular factorization $V = S.S^T$, and the one with positive diagonal elements is given by the following algorithm: For j = 1,...,d-1 (d= dim V), recursively cycle through the following ordered set of equations:

$$S(j,j) \quad = \quad V(j,j)^{\frac{1}{2}}$$

$$S(k,j) = V(k,j)/S(j,j) \quad , \quad k = j+1, .., d.$$

$$V(i,k) \longleftarrow V(i,k) - S(i,j).S(k,j) \quad , \quad k = j+1, ., d; \; i = k, ., d$$

and then $S(d,d) = V(d,d)$

Proof: (see Bierman,G.J.(1977) for instance)

Comments: i) An upper triangular factorization follows a similar algorithm , changing only indices order.

ii) The Cholesky decomposition algorithm can also be phrased in terms of the Sweep Operator - Goodnight(1979).

B) The Jacob Method and related techniques for Eigen-Structure computing

Definition: A dxd matrix V is said to have an eigen-vector $\underline{x}$ and corresponding eigenvalue λ if $V.\underline{x} = \lambda.\underline{x}$. Multiples of $\underline{x}$ are not considered distinct eigen-vectors and the zero vector is not considered to be an eigen-vector at all.

Basic Properties:

i)Existence: The associated characteristic equation $|V - \lambda.I| = 0$, derived from the definition , is a d^{th} degree polynomial in λ whose roots are the eigenvalues , that proves that there exist

always d (not necessarily distinct) eigenvalues. Consequently, by the definition , there are d corresponding eigen-vectors (not necessarily distincts).

ii) Symmetry & diagonalization: If V is symmetric , the eigenvalues of V will be real. Also, if they are distinct , the eigen-vectors $\underline{x}$ will be orthogonal to each other , forming an orthogonal matrix X. Then, by definition , $V.X = X.diag(\lambda_1,..,\lambda_d)$, or equivalently , $X^{-1}.V.X = diag(\lambda_1,..,\lambda_d)$, which is a special (similarity) transformation called diagonalization.

iii) Similarity transformations: In general , for some transformation matrix X , the application $V \longrightarrow X^{-1}.V.X$ is called a similarity transformation of the matrix V. They play a crucial role in the computation of eigenvalues because they leave the eigenvalues unchanged , since

$$|X^{-1}.V.X - \lambda.I| = |X^{-1}(V - \lambda.I).X| = |V - \lambda.I|$$

In general, the strategy of virtually all modern eigensystem routines is to nudge the matrix V towards diagonal form by a sequence of similarity transformations $V \longrightarrow V_1 = X_1^{-1}.V.X_1$; $V_1 \longrightarrow V_2 = X_2^{-1}.V_1.X_2$ ..etc , till we get $V_k = diag(\lambda_1,..,\lambda_d)$ for a certain k. And, the eigen-vectors are the columns of the accumulated transformation $X = X_1.X_2....X_k$.

Jacobi Transformations : The simplest technique for diagonalization of a symmetric matrix consists in a sequence of similarity transformations based on the Jacob rotation matrix,

$$P_{pq} = \begin{pmatrix} 1 & . & . & . & . \\ . & c & . & s & . \\ . & . & 1 & . & . \\ . & -s & . & c & . \\ . & . & . & . & 1 \end{pmatrix}$$

where all the diagonal elements are unity , except for the two elements c in rows (and columns) p and q. All off-diagonal elements are zero , except the two elements s and -s. The numbers c and s are the cosine and sine of a rotation angle ϕ , where these elements are chosen (at each time) in order to make one of the off-diagonal elements of the transformed matrix equal zero. The original Jacob's algorithm searched the whole upper triangle at each stage and set the largest off-diagonal element to zero. A better strategy however, used nowadays, is the cyclic Jacob method, where one annihilates elements in strick order.

Comment : It is important to mention however that when d is large (around 10 or more), the Jacobi method became inefficient , and is recommended to reduce V first to a tridiagonal form (using Household transformations) and then use some iterative or factorization method (Jacobi , QL algorithm, etc) to complete the diagonalization. The Household algorithm reduces a dxd

symmetric matrix V to tridiagonal form by d-2 orthogonal transformations , each one annihilating part of a whole column and corresponding row. The basic ingredient is the orthogonal matrix $P = 1 - 2.\underline{w}.\underline{w}^T, \quad |\underline{w}|^2 = 1$. It is easy to show that P acts on a given vector $\underline{x}$ to zero all its elements except the first one. The implementation details of this method can be found for instance in Press,W.H.at all(1986).

FIG. 6.1 PRICES OF WHEAT IN EUROPEAN COUNTRIES

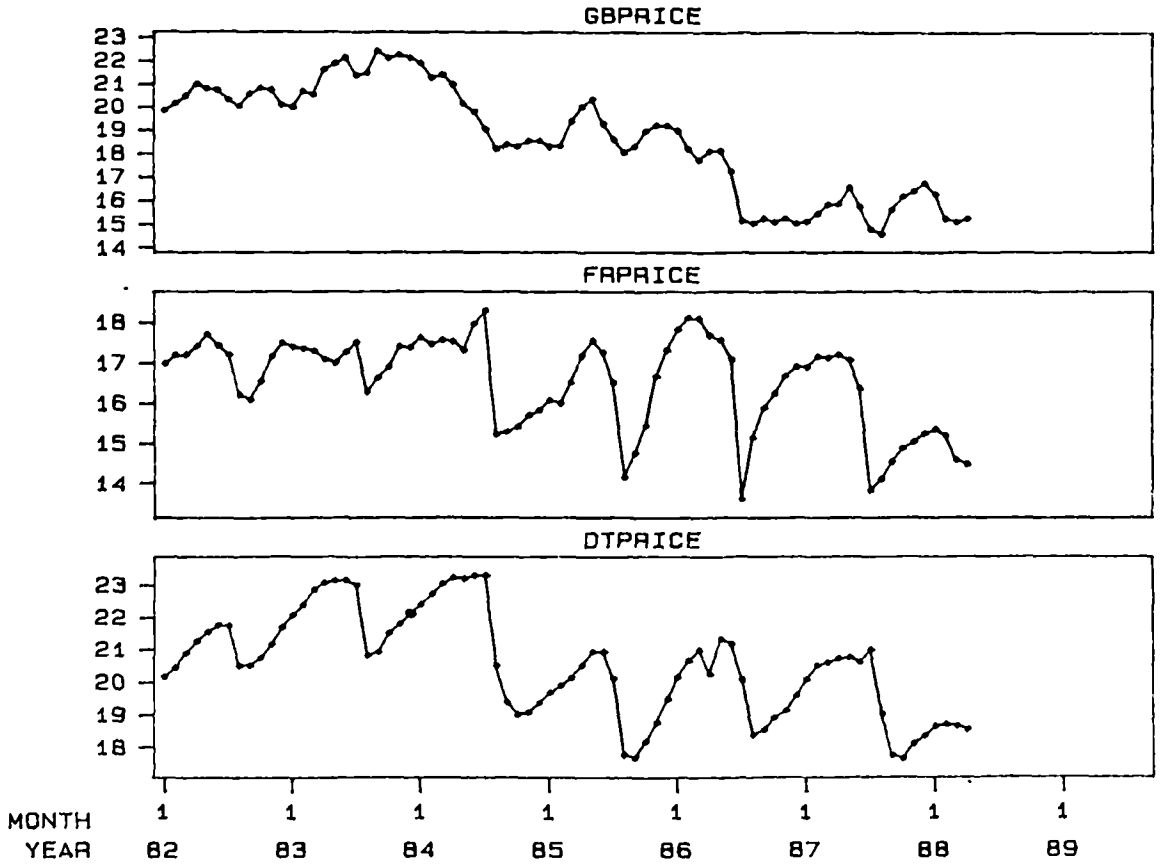


FIG. 6.2 ONE STEP AHEAD FORECASTING RESIDUALS

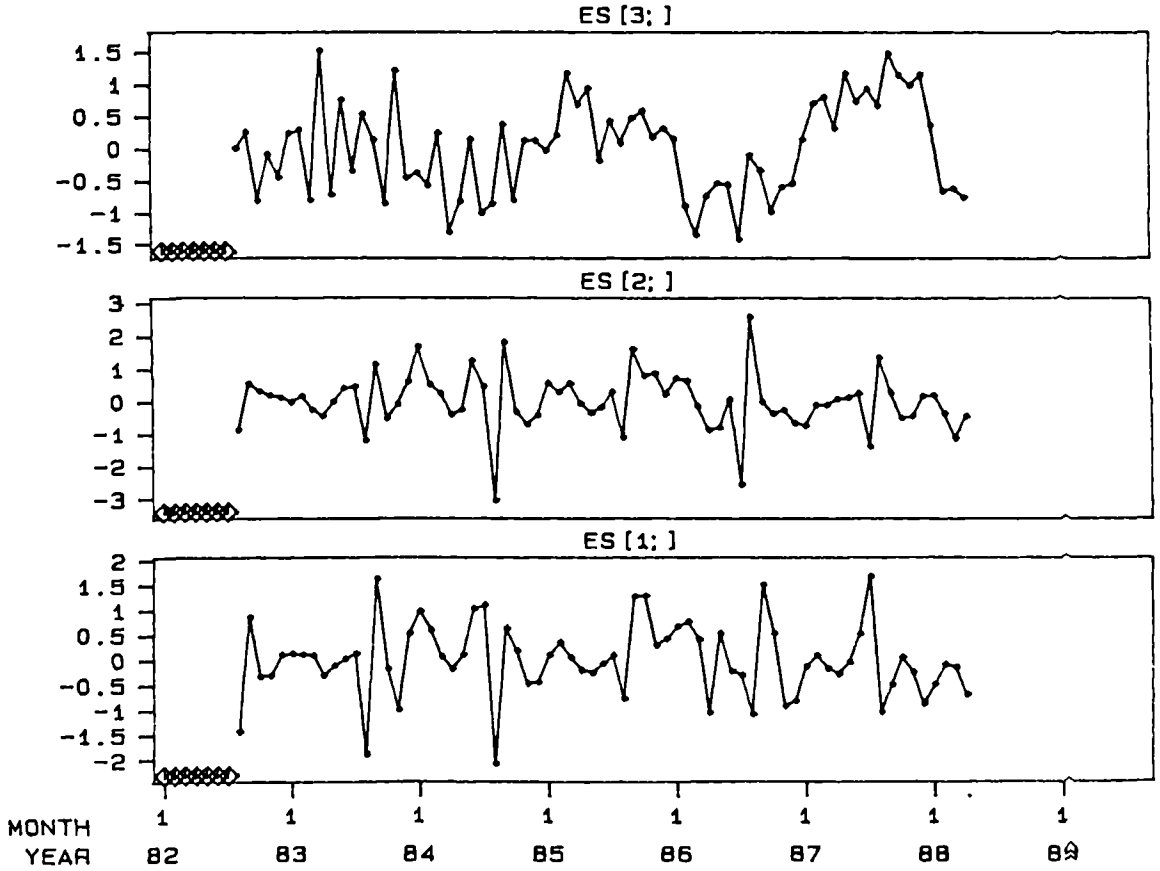


FIG. 6.3 CORRELATION BETWEEN THE COMPONENT SERIES

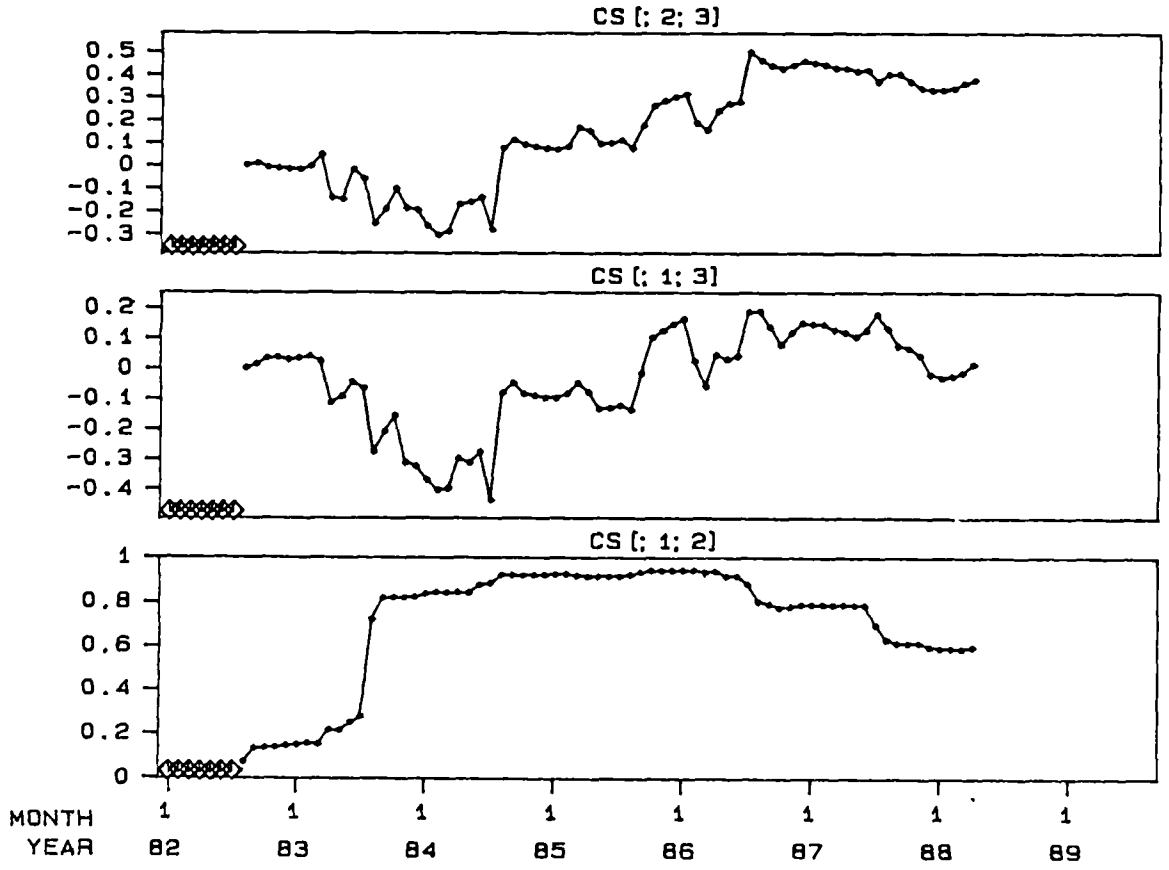


fig 6.4 PERIODOGRAM / DTPRICE SERIES

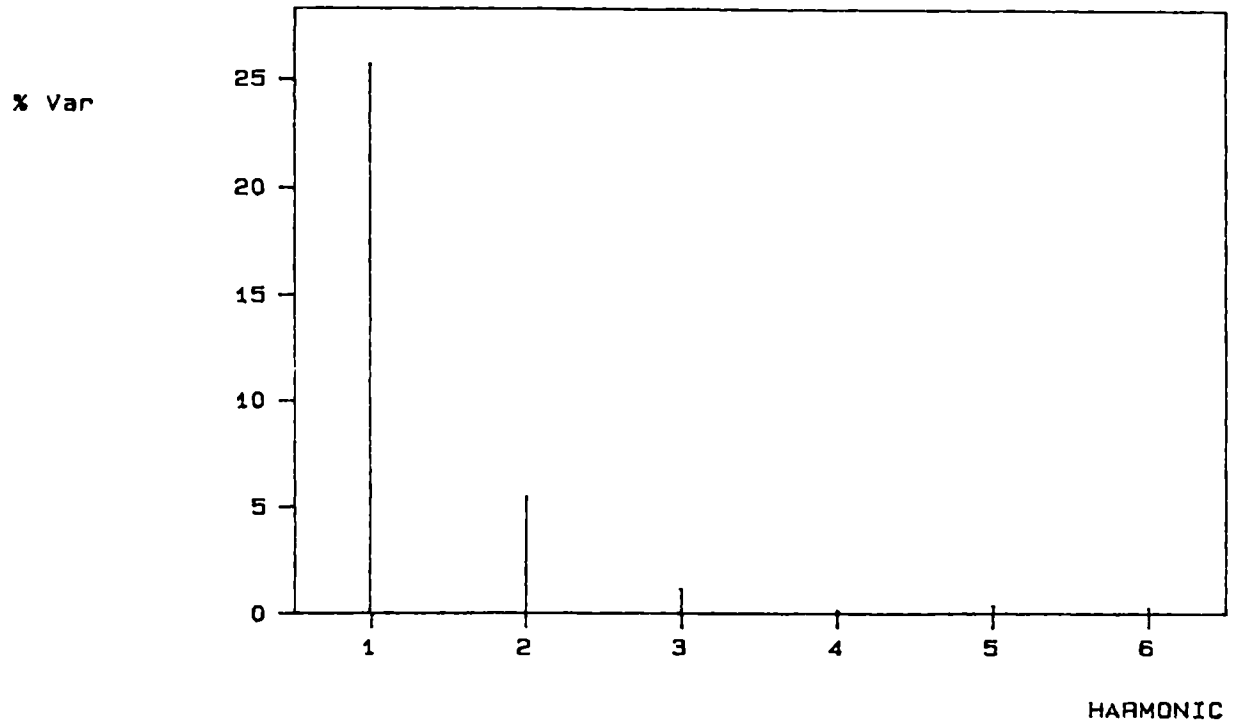


fig 6.5 PERIODOGRAM / FAPAICE SERIES

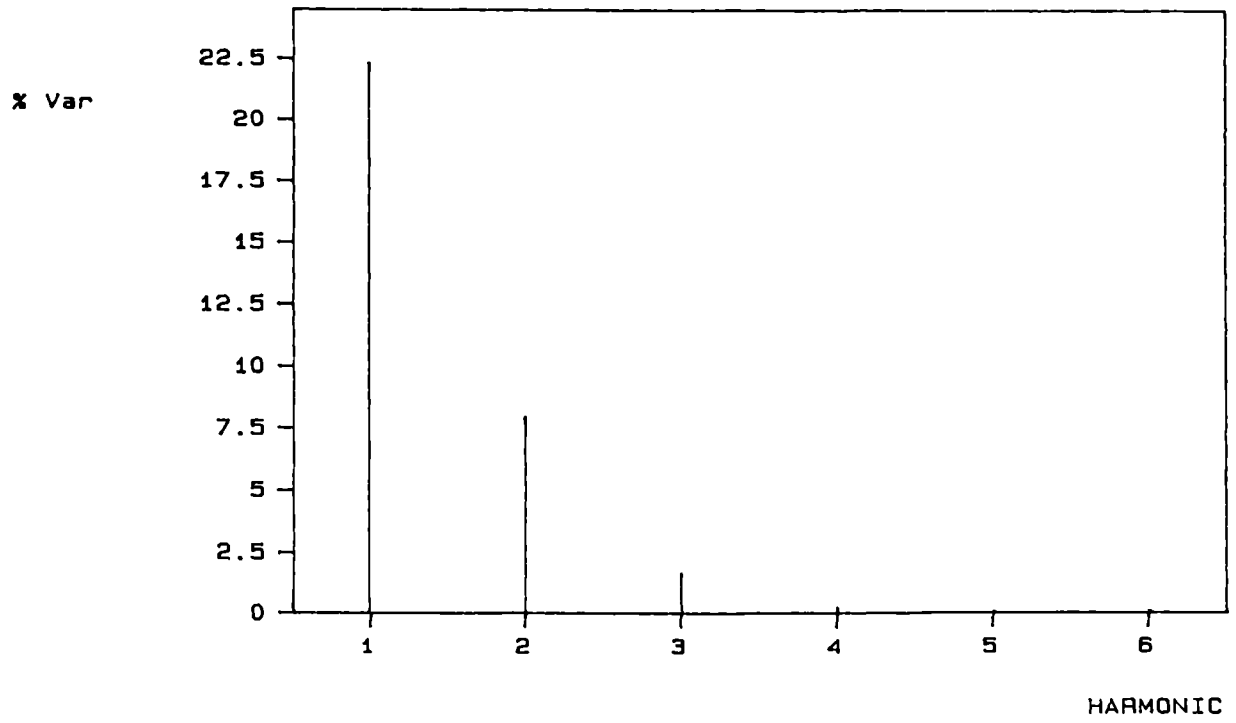
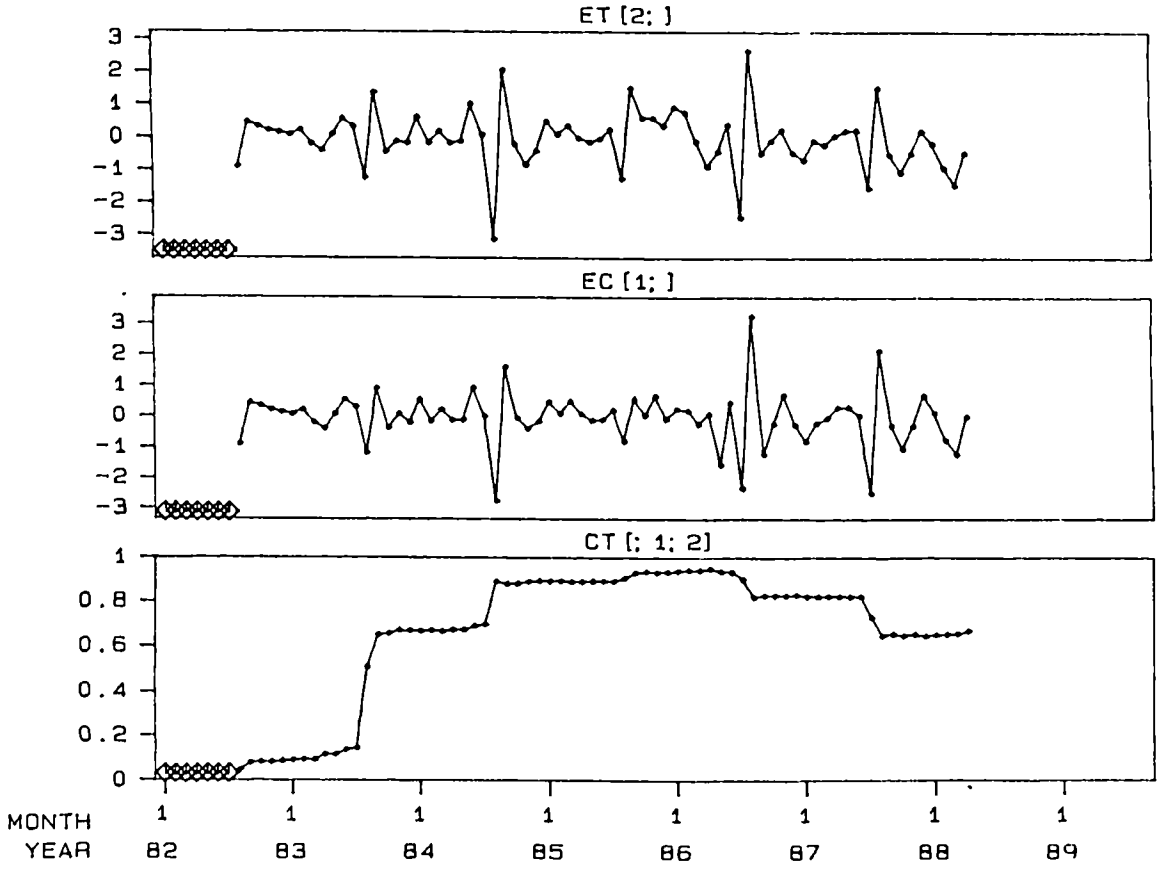


FIG 6.7 CONDITIONAL RESIDUALS FOR THE FARRICE SERIES



CHAPTER 7

SOME ALTERNATIVE TECHNIQUES FOR DLM ANALYSIS

7.1 - Introduction :

As we have seen in chapter 5 , fully closed bayesian procedures for estimating $\underline{\theta}_t$ and V in the multivariate D.L.M. model are available only for particular cases and an efficient estimation algorithm for the general case is far from trivial. However, alternative approximation techniques are possible , and we compare here the new methodology proposed in the last chapter , with some of these alternative techniques .

The more important of these alternative procedures for multivariate DLM analysis consists in obtaining an approximated inverse-Wishart marginal posterior distribution for V_t based on matrix Taylor series expansions - West (1982), and a marginal posterior distribution for $\underline{\theta}_t$ based on the Robust Filter - Masreliez (1975) & West (1981) . This procedure will be referred to as marginal approximation approach and is presented briefly in the next section of this chapter , where it is compared with the new methods of last chapter in relation to some aspects .

In the following section , some alternative approximation techniques proposed in the engineering literature such as sequential likelihood methods - Maybeck(1982) , are presented and discussed comparatively . Also , the possibility of using Linear Bayes tools for joint estimation of $\underline{\theta}_t$ and V is discussed in the last section of this chapter , as well as the drawbacks associated with such approach . Finally , some technical results are presented at the end of this chapter as Appendices as well as some graphical results .

7.2 The Marginal Approximation approach

This method is based on the Robust Filter , introduced by Masreliez(1975) and used by West(1982) as an approximation for the marginal posterior distribution for the process parameter $\underline{\theta}_t$ in a multivariate (normal) DLM context . In addition , an inverse-Wishart approximation to the marginal posterior distribution for V is presented by West(1982) , based on Taylor series truncation of the exact posterior distribution .

7.2.1 - A marginal posterior distribution for V

Considering the usual notation of the last two chapters , the likelihood for V can be written

as,

$$p(\underline{y}_t/V, D_{t-1}) \propto |Q_t|^{-\frac{1}{2}}.exp[-\frac{1}{2}(\underline{y}_t - F_t^T.\underline{a}_t)^T.Q_t^{-1}.(\underline{y}_t - F_t^T.\underline{a}_t)]$$

where $Q_t = F^T.R_t.F + V$. Equivalently, using properties of trace of products,

$$p(\underline{y}_t/V, D_{t-1}) \propto |Q_t|^{-\frac{1}{2}}.exp[-\frac{1}{2}.trQ_t^{-1}.\underline{e}_t.\underline{e}_t^T] \tag{7.1}$$

where $\underline{e}_t$ is the one-step-ahead forecasting error.

Also, using the inverse-Wishart distribution as prior for V or, equivalently, a Wishart distribution as prior for V^{-1} ,

$$p(V^{-1}/D_{t-1}) \propto |V^{-1}|^{\frac{n_{t-1}-d-1}{2}}.exp[-\frac{1}{2}.trV^{-1}.S_{t-1}] \tag{7.2}$$

we get by Bayes rule, after some algebra, the following marginal posterior distribution:

$$p(V^{-1}/D_t) \propto |V^{-1}|^{\frac{n_t-d-1}{2}}.exp[-\frac{1}{2}.tr(ln|V^{-1}| + ln\ Q_t| + V^{-1}.S_{t-1} + Q_t^{-1}.\underline{e}_t.\underline{e}_t^T)] \tag{7.3}$$

Now, approximating the exponent of (7.3) by Taylor series expansion as a linear function of the precision matrix $P = V^{-1}$ in order to have the same Wishart structure as the prior in (7.2), we get, after some algebra, the following updating equations: (details in Appendix 7.1a)

$$d_t = d_{t-1} + Q_t^{-1}.V_{t-1}.[\underline{e}_t.\underline{e}_t^T + (Q_t - V_{t-1}).V_{t-1}^{-1}.Q_t].V_{t-1}.Q_t^{-1} \tag{7.4}$$

and also, naturally,

$$n_t = n_{t-1} + 1 \tag{7.5}$$

$$V_t = E[V/D_t] = (n_t - d)^{-1}.d_t \tag{7.6}$$

7.2.2 - Marginal Posterior for $\underline{\theta}_t$

Considering $p(\underline{\theta}_t/D_{t-1}) \sim N(\underline{a}_t\ ,\ R_t\ )$ as prior distribution for $\underline{\theta}_t$, since the observational variance V is unknown , we can write the predictive distribution for the observations $\underline{y}_t$ as a multivariate t distribution,

$$p(\underline{y}_t/D_{t-1}) \propto [n_{t-1} + (\underline{y}_t - F_t.\underline{a}_t)^T.Q_t^{-1}.(\underline{y}_t - F_t.\underline{a}_t)]^{-\frac{n_{t-1}+1}{2}} \tag{7.7}$$

or, equivalently, in short form ,

$$p(\underline{e}_t) \propto (n_{t-1} + \underline{e}_t^T.Q_t^{-1}.\underline{e}_t)^{-\frac{n_{t-1}+1}{2}} \tag{7.7a}$$

where $\underline{e}_t = \underline{y}_t - F_t.\underline{a}_t$ and $Q_t = F_t^T.R_t.F_t + V_{t-1}$.

An approximation to the sequential updating of $\underline{\theta}_t$ can be obtained through the Robust Filter equations (details can be found in the Appendix 7.1b) as follows : $(\underline{\theta}_t/D_t) \sim N(\underline{m}_t, C_t)$,

$$\underline{m}_t = \underline{a}_t + R_t.F_t^T.g(\underline{y}_t) \tag{7.8}$$

$$C_t = R_t - R_t.F_t^T.H(\underline{y}_t).F_t.R_t \tag{7.9}$$

where:

$$g(\underline{y}_t) = -\frac{\partial}{\partial \underline{y}_t} lnp(\underline{y}_t) = \frac{Q_t^{-1}}{\alpha_t}.\underline{e}_t \tag{7.10}$$

$$H(\underline{y}_t) = \frac{\partial}{\partial \underline{y}_t} g(\underline{y}_t) = \frac{Q_t^{-1}}{\alpha_t} \tag{7.11}$$

with:

$$\alpha_t = (1 + n_{t-1})^{-1}.(n_{t-1} + \underline{e}_t^T.Q_t^{-1}.\underline{e}_t) \tag{7.12}$$

where the derivatives in (7.10)-(7.11) are easily obtained from (7.7) .

7.2.3 - Methods Assessment

It is interesting to remember that one key feature a general joint estimator for $\underline{\theta}_t$ and V should possess (and not present in the common component model of chapter 5) is that the posterior mean for $\underline{\theta}_t$ should depend on the observational noise variance parameters. Or , in other words , the mean updating for $\underline{\theta}_t$ should depend on the uncertainty about V , as mentioned before in chapter 5 .

In the case of the marginal approximation approach , the posterior mean for $\underline{\theta}_t$, expressed by equation (7.8) , is a function of the precision matrix parameters through the equation (7.10) for g(.) and the sort of limitations mentioned before doesn't occur here. In fact , the uncertainty about the observational variance was introduced into the Robust filter equations through the specification of a multivariate t distribution as the predictive distribution for $\underline{y}_t$.

Another important property , also present in the new methods of the last chapter , is the coincidence with conjugate analysis results under common components conditions . In order to investigate if this key property is valid or not for the marginal approximation approach we present here the following analysis :

i) Using the Robust filter equations , the marginal posterior distribution for $\underline{\theta}_t$ will be approximated by : $(\underline{\theta}_t/D_t) \sim N(\underline{m}_t, \mathbf{C}_t)$ where

$$\underline{m}_t = \underline{a}_t + \frac{\mathbf{A}_t}{\alpha_t}.(\underline{y}_t - \underline{f}_t) \tag{7.13}$$

$$\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t.\frac{\mathbf{Q}_t}{\alpha_t}.\mathbf{A}_t^T \tag{7.14}$$

where $\mathbf{A}_t = \mathbf{R}_t.\mathbf{F}^T.\mathbf{Q}_t^{-1}$. Under common components conditions , using (7.13)-(7.14) we have

$$\underline{m}_t = \underline{a}_t + \frac{(I \otimes A_t)}{\alpha_t}.(\underline{y}_t - \underline{f}_t) \tag{7.15}$$

$$\mathbf{C}_t = V_{t-1} \otimes \left(R_t - A_t.\frac{Q_t}{\alpha_t}.A_t^T\right) \qquad \text{where} \qquad A_t = R_t.F^T.Q_t^{-1} \tag{7.16}$$

and , since $\alpha_t = (n_{t-1}+1)^{-1}.(n_{t-1} + \underline{e}_t^T.\mathbf{Q}_t^{-1}.\underline{e}_t)$, this posterior distribution approach conjugate analysis results only for a number of d.f. such that $n_{t-1} \gg \underline{e}_t^T.\mathbf{Q}_t^{-1}.\underline{e}_t$, when α_t approach 1 . Or , in other words , there is convergence to conjugate analysis results , since the Student t filter converge to the normal filter as the d.f. increases - [West(1982),chapter 4] . See also Appendix - 7.1b .

ii) From the inverse-Wishart approximation given by the equations (7.3)-(7.5) , we have that $p(V^{-1}/D_{t-1}) = W(d_t, n_t)$, where

$$d_t = d_{t-1} + \mathbf{Q}_t^{-1}.V_{t-1}.\left(\underline{e}_t.\underline{e}_t^T + Var(\mathbf{F}.\underline{\theta}_t/D_{t-1}).V_{t-1}^{-1}.\mathbf{Q}_t\right).V_{t-1}.\mathbf{Q}_t^{-1} \tag{7.17}$$

But , under common components conditions , we have

$$\mathbf{Q}_t^{-1}.V_{t-1} = (V_{t-1} \otimes Q_t)^{-1}.V_{t-1} = Q_t^{-1}.I \qquad \text{where } Q_t^{-1} \text{ is scalar} \tag{7.17a}$$

$$Var(\mathbf{F}.\underline{\theta}_t/D_{t-1}) = (I \otimes F^T).(V_{t-1} \otimes R_t).(I \otimes F) = \lambda.V_{t-1} \qquad \text{where} \quad \lambda = F^T.R_t.F \tag{7.17b}$$

Then , the equation (7.17) under common components conditions can be rewritten as $d_t = d_{t-1} + \Delta_t$ where

$$\Delta_t = \beta_t.(\underline{e}_t\underline{e}_t^T.Q_t^{-1}) + (1-\beta_t).V_{t-1} \qquad \text{with} \qquad \beta_t = Q_t^{-1} = \frac{1}{1+\lambda} \tag{7.18}$$

and since $0 < \beta_t < 1$ ($\lambda > 0$ because there is uncertainty about $\underline{\theta}_t$) , there is no coincidence with the increment $\underline{e}_t.\underline{e}_t^T.Q_t^{-1}$ for the CCM (for details about this model see chapter 5) . When $\underline{\theta}_t = \underline{\theta}$ is not dynamic , then λ converge to zero , and consequently β_t converge to 1 ,

and we have convergence to conjugate analysis results in the variance updating . In general , when $\underline{\theta}_t$ is dynamic we have also consistency in the estimation of V using (7.17) , as pointed out by West(1982) . However , for finite samples , as the increment Δ_t is less dependent on the new data through the error term than in the CCM (where β_t is 1) , and consequently more conservative , we can not expect a variance learning process as quick as in the conjugate analysis . Consequently , we would expect a better variance learning process with the methods of last chapter than with the methods of this section , as confirmed through the use of simulated series .

Now , in order to complement the above analysis of the Marginal Approximation Approach - MAA , relative to the new methods of chapter 6 , we consider here a brief experiment with simulated data . The comparative performance of both methods , specially in relation to variances and correlation coefficients estimation are desired .

The simulated data consists of a bivariate time series following a linear growth model , where each series is described by a level parameter and a slope (growth) parameter , both changing slowly in time and related to the observations through the DLM structure . The series have 75 observations each and they were generated with observational variances of 4 and 10 respectivelly and a correlation coefficient between noise series of 0.5 . A time plot of this data , labeled as Series1 and Series2 , is shown in figure 7.1 and the vector series itself is presented in Tab 7.1 .

Considering this data and the use of non-informative initial priors (large values for the parameters representing uncertainty : $C_0 = 1000.I \,\&\, n_0^{-1} = 1000$) the marginal approximation method was applied and , in particular , the time sequences $\{\alpha_t\}$ and $\{\beta_t\}$ were monitored , with the results shown in the figs. 7.8 - 7.9 . Also , using the same initial priors , one-step-ahead forecasting errors (figs. 7.2 - 7.3) , observational noise variances (figs. 7.4-7.5) and intra-series correlations (figs. 7.6-7.7) were monitored comparatively for both methods . The notation used in these graphics is as follows : EW , VW & CW denote respectively the Errors , Variances and Correlations (as explained in the respective titles) for the MAA method (or West method) and similarly , ES , VS & CS represent the same quantities for the method of chapter 6 . For instance , ES[1;] represents the errors (in the one-step-ahead forecasting) for the Series1 using the new method of chapter 6 ; VW[;2;2] represents the observational variance for the Series2 , estimated via the MAA method , and so on .

Based on the analytical and numerical results above mentioned , we have the following comments about the methods individual and comparative performance :

i) Not only the MAA updating for $\underline{\theta}_t$ converge to the CCM updating as the d.f. increases as seen in i) but also this convergence is very quick (see fig. 7.8) . As a consequence , both methods in general will behave very similarly for location estimation or prediction (see figs. 7.2-7.3 for 1-step-ahead forecasting error comparative performance) .

ii) The new method updating for V is by far better than the MAA updating . In the case of correlations for instance , as we can see from the fig.7.6-7.7 , the MAA algorithm is not able to reach the true correlation value 0.5 at the end of 75 observations while the new algorithm does it at the beginning . (both methods are initialized with zero correlation and a relatively vague prior information in the other parameters ; the initial d.f. ranges from 0.01 in fig.7.6 to 1 in fig.7.7) . For variance estimation however , the MAA algorithm performs not totally bad (see fig.7.4-7.5) but still not as good as the new algorithm .

iii) As we have seen before , the new algorithm of chapter 6 (under common components conditions and using a standard diagonal discount squeme) does not coincide with the conjugate analysis algorithm results , and the coincidence is obtained only using an alternative constant discount scheme . However , in most practical situations , both discount schemes give results almost undistinguishable . That is the case for instance , in the example given by the figure 7.10 , where a discount factor of 0.98 was used .

7.3 - Likelihood Approximation Approach :

In order to investigate other possible alternative forms of model analysis in a context as wide as possible , we consider also here some approximated procedures based on maximum likelihood ideas . This method is one of the more accepted and more substantiated tools in the engineering literature about system identification . One of the reasons perhaps is because some of its properties make it attractive for certain applications . For details about these properties , see for instance , Rao,C.R.[1973] . However , one big disadvantage to maximun likelihood estimation (apart from the difficulties and limitations shown in this special application) is the lack of theoretical knowledge about the behavior of the estimates for small sample sizes .

In order to make explicit the difficulties and limitations of such likelihood procedures in problems of sequential DLM estimation we present the basic ideas used in the development of

problems of sequential DLM estimation we present the basic ideas used in the development of these methods . The construction of such methods involve two basic steps : the formulation of the likelihood equations (exact or approximated) and the solution of such equations . These two steps are presented briefly in the next couple of sub-sections and it is followed by some critic discussion .

7.3.1 -The Likelihood Equations :

By repeated use of probability laws , we can write the likelihood for the observational variance V in the multivariate DLM as :

$$\begin{aligned} p(\underline{\theta}_t, D_t/V) &= p(\underline{\theta}_t/D_t,V).p(D_t/V) \\ &= p(\underline{\theta}_t/D_t,V).p(\underline{y}_t/D_{t-1},V).p(D_{t-1}/V) \\ &= \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &= p(\underline{\theta}_t/D_t,V).\prod_{j=1}^{t} p(\underline{y}_j/D_{j-1},V) \end{aligned} \tag{7.19}$$

Or , equivalently , in terms of log-densities (log-likelihoods), we have :

$$ln\, p(\underline{\theta}_t, D_t/V) = ln\, p(\underline{\theta}_t/D_t,V) + \sum_{j=1}^{t} ln\, p(\underline{y}_j/D_{j-1},V) \tag{7.20}$$

For a multivariate normal DLM , each of the log-densities in (7.20) can be written out explicitly as :

$$ln\, p(\underline{\theta}_t/D_t,V) = const. - \frac{1}{2}.ln|C_t| - \frac{1}{2}.(\underline{\theta}_t - \underline{m}_t)^T.C_t^{-1}.(\underline{\theta}_t - \underline{m}_t) \tag{7.20a}$$

$$ln\, p(\underline{y}_j/D_{j-1},V) = const. - \frac{1}{2}.ln|Q_t| - \frac{1}{2}.(\underline{y}_j F^T.\underline{a}_j)^T.Q_j^{-1}.(\underline{y}_j - F^T.\underline{a}_j) \tag{7.20b}$$

where $Q_j = F^T.R_j.F + V$. Note that all these expressions are implicitly functions of V . Now , by subst. the expressions (7.20a)-(7.20b) in the log-likelihood function , we have :

$$\begin{aligned} ln\, p(\underline{\theta}_t, D_t/V) = const. - \frac{1}{2}.ln|C_t| - \frac{1}{2}.\sum_{j=1}^{t} ln|Q_t| - \frac{1}{2}(\underline{\theta}_t - \underline{m}_t)^T.C_t^{-1}.(\underline{\theta}_t - \underline{m}_t) \\ - \frac{1}{2}.\sum_{j=1}^{t}(\underline{y}_j - F^T.\underline{a}_j)^T.Q_j^{-1}.(\underline{y}_j - F^T.\underline{a}_j) \end{aligned} \tag{7.21}$$

And , the likelihood equations will be given by its partial derivatives :

$$\left.\frac{\partial}{\partial \underline{\theta}_t} ln\, p(\underline{\theta}_t, D_t/V)\right|_{V=\hat{V}_t} = \underline{0} \tag{7.21a}$$

$$\left.\frac{\partial}{\partial V} ln\, p(\underline{\theta}_t, D_t/V)\right|_{\underline{\theta}_t=\hat{\underline{\theta}}_t} = \underline{0} \tag{7.21b}$$

Using (7.21) to develop the equation (7.21a) , we get

$$-(\underline{\theta}_t - \underline{m}_t).C_t^{-1}\big|_{V=\hat{V}_t} = \underline{0} \qquad \longrightarrow \qquad \underline{\hat{\theta}}_t = \underline{m}_t\Big|_{V=\hat{V}_t} \tag{7.22}$$

Or , the maximum likelihood estimate $\underline{\hat{\theta}}_t$ of $\underline{\theta}_t$ is given by the standard posterior mean $\underline{m}_t$, where $\hat{V}_t$ (the ML estimate of V) replaces V through its expression ($Q = F^T.R.F + V$) .

Also , from (7.21b) , after some calculations and simplifications (see Appendix 7.2 for details) we get the following approximated likelihood equation for the elements of V :

$$\sum_{j=t-N+1}^{t} tr\{[Q_j^{-1} - Q_j^{-1}.\underline{u}_j.\underline{u}_j^T.Q_j^{-1}].\frac{\partial Q_j}{\partial \nu_k}\}\Big|_{\underline{\theta}_t=\underline{m}_t} = \underline{0} \tag{7.23}$$

where $\underline{u}_j = \underline{y}_j - F.\underline{a}_j$; ν_k is the k^{th} element of vec V , and N is such that V is supposed essentially constant over N periods of time .

7.3.2 - L.E. Approximated Solution :

Now , since $Q_j = F^T.R_j.F + V$, the partial derivative of Q_j that appears in (7.23) is given by :

$$\frac{\partial Q_j}{\partial \nu_k} = F^T.\frac{\partial R_j}{\partial \nu_k}.F + \frac{\partial V}{\partial \nu_k} \cong I \tag{7.23a}$$

because we neglect $\frac{\partial R_j}{\partial \nu_k}$, since the estimation of $\underline{\theta}_t$ is robust in relation to variations in V . Now , rewritting the equation (7.23) , we get

$$\sum_{j=t-N+1}^{t} tr\{Q_j^{-1}.[Q_j - \underline{u}_j.\underline{u}_j^T].Q_j^{-1}\}\Big|_{\underline{\theta}_t=\underline{m}_t} = 0 \tag{7.23b}$$

which would be satisfied for all j if the term between brackets is zero . Also , if the estimation process is essentially time invariant over the most recent N steps , i.e. , $Q_j^{-1} \cong const.$ over these steps , then an estimate $\hat{V}_t$ of V at time t can be obtained from ,

$$\sum_{j=t-N+1}^{t} [F^T.R_j.F + V - \underline{u}_j.\underline{u}_j^T] = 0 \qquad \text{what gives ,}$$

$$\hat{V}_t = \frac{1}{N}\sum_{j=t-N+1}^{t} [\underline{u}_j.\underline{u}_j^T - F^T.R_j.F] \tag{7.24}$$

$$\hat{V}_t = \frac{1}{N}\sum_{j=t-N+1}^{t} \underline{u}_j.\underline{u}_j^T - F^T.R_t.F \tag{7.24a}$$

It happens however that these expressions , also known as Covariance matching estimator - Mehra(1972) , Chin(1979) - can lead to an estimate of V that is not positive definite . In order to avoid these difficulties , a better estimate can be obtained by verifying the following relation :

$$Q_j^{-1}.\underline{u}_j^* = V^{-1}.(\underline{y}_j - F^T.\underline{m}_j) \tag{7.25}$$

where $\underline{u}_j^* = \underline{y}_j - F.\underline{m}_j$ is a sort of posterior residual. Now , using (7.25) into equation (7.23b) , we have :

$$\hat{V}_t = \frac{1}{N} \sum_{j=t-N+1}^{t} [\underline{u}_j^*.\underline{u}_j^{*T} + F^T.C_j.F] \tag{7.26}$$

$$\hat{V}_t = \frac{1}{N} \sum_{j=t-N+1}^{t} \underline{u}_j^*.\underline{u}_j^{*T} + F^T.C_t.F \tag{7.26a}$$

7.3.3 - Method assessment and discussion :

Although the adaptation of maximum likelihood techniques for using in a sequential estimation context is theoreticaly sounded , its practical implementation requires certain approximations , resulting in equations such as (7.24)-(7.24a) which are equivalent to some ad hoc procedures (Covariance matching techniques) present in the engineering literature .

In particular , the use of posterior information about $\underline{\theta}_t$ in the estimation of V at time t introduced through equation (7.25) improves the method considerably but the variance estimation is still an off-line procedure . In fact , we can consider the whole method as a kind of two-stage procedure : given V , the process parameter $\underline{\theta}_t$ is estimated in a standard Bayesian way ; after that , V is estimated using an auxiliary relation (equation 7.26 or 7.26a) . This method presents the same kind of drawback as the ones mentioned in section 5.2.3 and a more deeper analysis follows .

From equations (7.26) or (7.26a) we can see that the uncertainty about $\underline{\theta}_t$ expressed through C_t is present in the estimation equation for V_t , which characterizes joint estimation and is a desirable property . However , the uncertainty about V is neglected in the estimation of $\underline{\theta}_t$, which is supposed (as an approximation) to be robust in relation to variations in V . Or , in other words , the approximation uses the standard Kalman Filter equations for the estimation of the process parameter $\underline{\theta}_t$, which is a limitation . It is important to remember that this sort of limitation does not occur with our new propposed methods of chapter 6 or even with the marginal approximation approach of this chapter .

Also , one very important practical aspect to be considered in the implementation of equation (7.26) or (7.26a) is the choice of a suitable value for N . This constant is a sort of smooth parameter in the sense that a sequential estimation of V with only small irregularities will demand a large value for N . In fact , we would expect from a good estimator for V , a smooth sequence of estimates , i.e., a sequence of estimates with only small fluctuations after a certain number of observations have been processed .

Then , it is clear that we need to use a large value for N , but this will introduce at least two inconveniences . First , a large N will be computationally undesirable because of the quantities that need to be stored , mainly in (7.26) . Second , there is an initialization problem in order to implement the sum in equation (7.26) or (7.26a) and a large number of extra initial observations will be necessary when N is large .

Also , experiments with real and simulated data have shown that is very difficult to choose a value for N that could conciliate these problems , which makes this method very limited in practice .

7.4 - The Linear Bayes Approach :

From the basic theory about linear Bayes estimation , presented briefly in Appendix 7.3 , it should be possible in principle , to use such ideas for the joint estimation of $\underline{\theta}_t$ and V in a multivariate DLM context . In order to make Linear Bayes Estimation operational we need to set up two initial steps :

i) elements definition : the data (or, a sufficient statistic) $\underline{z}_t$ and the parameter $\underline{\pi}_t$ should be specified .

ii) prior moments : $E(\underline{z}_t)$; $V(\underline{z}_t)$ & $Cov(\underline{z}_t, \underline{\pi}_t)$ should be determined .

For the data (or , sufficient transformation) definition , since we are estimating not only $\underline{\theta}_t$ but also variances , the data or a sufficient statistic from the data should include not only the original observations but also its squares which in a vectorial context , can be represented by ,

$$\underline{z}_t = \begin{pmatrix} \underline{y}_t \\ \underline{y}_t \otimes \underline{y}_y \end{pmatrix} \qquad \text{where} \quad vec\, \underline{y}_t . \underline{y}_t^T = \underline{y}_t \otimes \underline{y}_t \tag{7.27}$$

is the standard column vectorization of the matrix $\underline{y}_t . \underline{y}_t^T$ and $\otimes$ is the usual direct Kronecker product operator . (alternatively , we could consider the more parsimonious short vectorization operator VEC , where there are no repeated elements) .

The natural parameter definition for such a multivariate model could be represented in principle by $\underline{\pi}_t = \begin{pmatrix} \underline{\theta}_t \\ \underline{\sigma}_t \end{pmatrix}$ where $\underline{\sigma}_t = VEC\, V_t$. However , with the introduction of $\underline{y}_t \otimes \underline{y}_t$ in $\underline{z}_t$, the calculation of $V(\underline{z}_t)$ and $Cov(\underline{z}_t, \underline{\pi}_t)$ will require the specification of quantities such as $Cov(\underline{\theta}_t \otimes \underline{\theta}_t\ ,\ \underline{\sigma}_t)$ which are not given directly from the model.

This sort of difficulty could in principle be overcome through one of the following two possibilities . The first one could be the introduction of a relation between the variance-covariance matrices $\mathrm{Var}(\underline{\theta}_t)$ and V_t such as , for instance , a proportionality relation $\mathrm{Var}(F.\underline{\theta}_t/V_t) \propto V_t$ (or , perhaps a relation such as $\mathrm{Var}(\underline{\theta}_t) = R_t \otimes V_t$, used in the common components model of chapter 5) .

In order to avoid such procedures , a second and more natural solution can be the introduction of $\underline{\alpha}_t = \underline{\theta}_t \otimes \underline{\theta}_t$ in the basic parametrization $\underline{\pi}_t$,

$$\underline{\pi}_t = \begin{pmatrix} \underline{\theta}_t \\ \underline{\alpha}_t \\ \underline{\sigma}_t \end{pmatrix} \qquad \text{where} \quad \underline{\alpha}_t = VEC\, \underline{\theta}_t.\underline{\theta}_t^T \quad \& \quad \underline{\sigma}_t = VEC\, V_t \tag{7.28}$$

Then , we have completed the basic definitions concerning $\underline{z}_t$ and $\underline{\pi}_t$, and the model can be defined formally as follows :

$$\text{i) obs. equation :} \quad \underline{y}_t = F.\underline{\theta}_t + \underline{v}_t \qquad \underline{v}_t \sim (\underline{0}, V) \tag{7.29}$$

$$\text{ii) system equation :} \quad \underline{\theta}_t = G.\underline{\theta}_{t-1} + \underline{w}_t \qquad \underline{w}_t \sim (\underline{0}, W_t) \tag{7.29a}$$

$$\text{iii) prior inf.:} \quad \begin{pmatrix} \underline{\theta}_t \\ \underline{\alpha}_t \\ \underline{\sigma}_t \end{pmatrix} / D_{t-1} \sim \left\{ \begin{pmatrix} \underline{m}_{t-1} \\ \underline{a}_{t-1} \\ \underline{s}_{t-1} \end{pmatrix} ; \begin{pmatrix} R_{t-1} & C_{t-1}^{12} & C_{t-1}^{13} \\ \cdot & U_{t-1} & C_{t-1}^{23} \\ \cdot & \cdot & P_{t-1} \end{pmatrix} \right\} \tag{7.30}$$

iv) assumption : $\underline{v}_t$ has same moments as the normal distribution up to 4^{th} order , and is independent of $\underline{\theta}_t$.

Now , the second step in the BLE implementation consists in the calculation of the following prior moments for $\underline{z}_t$:

$$\begin{aligned} i)\ E\underline{z}_t &= E_{\underline{\pi}_t} E(\underline{z}_t/\underline{\pi}_t) = E_{\underline{\pi}_t} \begin{pmatrix} F.\underline{\theta}_t \\ F^*.(\underline{\theta}_t \otimes \underline{\theta}_t) + \underline{\sigma}_t \end{pmatrix} \\ &= \begin{pmatrix} F.\underline{m}_{t-1} \\ F^*.(\underline{m}_{t-1} \otimes \underline{m}_{t-1} + vec\, R_{t-1}) + \underline{s}_{t-1} \end{pmatrix} \end{aligned} \tag{7.31}$$

where $F^* = F \otimes F$, since

$$E(\underline{\theta}_t \otimes \underline{\theta}_t) = E(vec\, \underline{\theta}_t.\underline{\theta}_t^T) = vec\, E(\underline{\theta}_t.\underline{\theta}_t^T) \qquad \text{and}$$

$$vec\,(\underline{m}_{t-1}.\underline{m}_{t-1}^T + R_{t-1}) = \underline{m}_{t-1} \otimes \underline{m}_{t-1} + vec\,R_{t-1}$$

$$ii) \quad V(\underline{z}_t) = Var\{E(\underline{z}_t / \underline{\pi}_t)\} + E\{Var(\underline{z}_t / \underline{\pi}_t)\}$$

where :

$$Var\,[E(\underline{z}_t / \underline{\pi}_t)] = \begin{pmatrix} F.R_{t-1}.F^T & F.C_{t-1}^{12}.F^{*T} + F.C_{t-1}^{13} \\ . & F^*.U_{t-1}.F^{*T} + P_{t-1} + 2.F^*.C_{t-1}^{23} \end{pmatrix} \tag{7.32a}$$

and also ,

$$E\,[Var(\underline{z}_t / \underline{\pi}_t)] = E\begin{pmatrix} V_t & L.[(F.\underline{\theta}_t) \otimes V_t] \\ . & L.[V_t \otimes V_t + (V_t \otimes F\underline{\theta}_t\underline{\theta}_t^T F^T).L] \end{pmatrix}$$

or equivalently,

$$= \begin{pmatrix} V_t & L(F \otimes I)(S_{t-1} \otimes \underline{m}_{t-1} + C_{t-1}^{13}) \\ . & L(S_{t-1} \otimes S_{t-1} + P_{t-1} + [S_{t-1} \otimes E(\underline{\psi}_t.\underline{\psi}_t^T) + C_{t-1}^{23}].L) \end{pmatrix} \tag{7.32b}$$

with $E(\underline{\psi}_t.\underline{\psi}_t^T) = F.[\underline{m}_{t-1}.\underline{m}_{t-1}^T + R_{t-1}].F^T$, where the expression for $Var(\underline{y}_t \otimes \underline{y}_t / \underline{\pi}_t)$ is developed in Appendix 7.2c .

$$iii) \qquad Cov(\underline{z}_t, \underline{\pi}_t) = \begin{pmatrix} F.R_{t-1} & F^*.C_{t-1}^{12} \\ F.C_{t-1}^{12} & F^*.U_{t-1} \\ F.C_{t-1}^{13} & F^*.C_{t-1}^{23} + P_{t-1} \end{pmatrix} \tag{7.33}$$

Putting together the prior moments for $\underline{\pi}_t$ (equation 7.30) and $\underline{z}_t$ (equations 7.31 & 7.32) , we have the following joint prior :

$$\begin{pmatrix} \underline{\pi}_t \\ \underline{z}_t \end{pmatrix} / D_{t-1} \sim \left\{ \begin{pmatrix} E\,\underline{\pi}_t \\ E\,\underline{z}_t \end{pmatrix} ; \begin{pmatrix} V(\underline{\pi}_t) & Cov(\underline{z}_t, \underline{\pi}_t) \\ . & V(\underline{z}_t) \end{pmatrix} \right\} \tag{7.34}$$

Then , the B.L.E equations will be given by :

$$E\,(\underline{\pi}_t / \underline{z}_t) = E\,\underline{\pi}_t + A_t.(\underline{z}_t - E\,\underline{z}_t) \tag{7.35}$$

$$Var(\underline{\pi}_t / \underline{z}_t) = Var(\underline{\pi}_t) - A_t.Var(\underline{z}_t).A_t^T \tag{7.36}$$

where $A_t = Cov(\underline{z}_t\,, \underline{\pi}_t).[Var(\underline{z}_t)]^{-1}$.

At this point , it is convenient to see that the introduction of $\underline{\alpha}_t$ in the model parametrization does not bring extra difficulties in terms of implementation such as initialization or time updating . In fact , these two implementational aspects can be handled as follows:

i) Initialization for $\underline{\alpha}_t = \underline{\theta}_t \otimes \underline{\theta}_t$: $\quad (t = 1)$

$\underline{\alpha}_t / D_{t-1} \sim (\underline{a}_{t-1}\,;\, U_{t-1})$ where :

$$\underline{a}_{t-1} = E(\underline{\theta}_t \otimes \underline{\theta}_t / D_{t-1}) = \underline{m}_{t-1} \otimes \underline{m}_{t-1} + vec\, R_{t-1}$$

$$U_{t-1} = V(\underline{\theta}_t \otimes \underline{\theta}_t / D_{t-1}) = L.[R_{t-1} \otimes R_{t-1} + (R_{t-1} \otimes \underline{m}_{t-1}.\underline{m}_{t-1}^T).L]$$

ii) Time-updating for $\underline{\alpha}_t = \underline{\theta}_t \otimes \underline{\theta}_t$: $(G \neq I \quad \& \quad W \neq 0)$

It is basically the same as the time-updating for $\underline{\theta}_t$, noticing that the system dynamics $\underline{\theta}_t = G.\underline{\theta}_{t-1}$ imply that $\underline{\alpha}_t = \mathbf{G}.\underline{\alpha}_{t-1}$ where $\mathbf{G} = G \otimes G$. Also , if we use a discount factor β for $\underline{\theta}_t$, we should use a discount factor β^2 for $\underline{\alpha}_t$.

At this point , we have a complete linear Bayes estimation methodology for joint estimation of $\underline{\theta}_t$ and V which is in principle , theoretically sounded and operationaly feasible . However , a deeper look at this procedure will show some remaining difficulties as explained below .

It is known in the context of DLM theory that $\underline{y}_t$ contain the sufficient information to estimate the parameter $\underline{\theta}_t$ and , in practice , we would like to have the estimation equation for $\underline{\theta}_t$ as a linear function of $\underline{y}_t$. It happens however that the coefficient of the quadratic term in $\underline{y}_t$ (element of A_t) is not zero and the method should be modified in order to attend this requirement .

One natural solution to overcome this drawback and guarantee linearity is to decouple $\underline{\theta}_t$ from the parametrization $\underline{\pi}_t$ keeping two separate sub-models , one for the location parameter $\underline{\theta}_t$ and other for , say $\underline{\gamma}_t = \begin{pmatrix} \underline{\alpha}_t \\ \underline{\sigma}_t \end{pmatrix}$. Also , as these two model components are interdependents , the interaction structure of covariances are updated separatedly as a third component.

Now , this new structured model is totally operational and gives updating equations for all quantities involved , but at least two drawbacks still remain . First , in order to take into account the uncertainty about V in the estimation of $\underline{\theta}_t$, the updating of the variance of $\underline{\theta}_t$ should be done indirectly through the variance of $\underline{\theta}_t \otimes \underline{\theta}_t$ which is correlated with $\underline{\sigma}_t$. In practice however , these variances and the ones obtained directly from the updating of $\underline{\theta}_t$ don't coincide (for coherence reasons they should) . The same problem occurs with the corresponding means . As a consequence , we do not get a positive estimate for V , which is one of the main drawbacks of Linear Bayes Estimation .

Appendix 7.1a - Matrix Taylor Series Expansion

Denoting by $f(P)$ the exponent in the expression (7.3) as a function of the precision matrix

$P = V^{-1}$, a first order matrix Taylor series expansion of f(.) at the point $P = P_{t-1}$ gives,

$$f(P) \cong f(P_{t-1}) + (P - P_{t-1}).\left.\frac{\partial f(P)}{\partial P}\right|_{P=P_{t-1}}$$

, where

$$f(P) = ln|P| + ln|Q_t(P)| + P.S_{t-1} + trQ_t^{-1}(P).\underline{e}_t.\underline{e}_t^T$$

In particular, we will use the following derivative results (see Press, pg 41) :

$$i) \quad \frac{\partial}{\partial P}|P| = P^{-1} = V$$

$$ii) \quad \frac{\partial}{\partial P} ln\,|Q(P)| = -V.Q_t^{-1}.V$$

$$iii) \frac{\partial}{\partial P} trQ_t^{-1}(P).\underline{e}_t.\underline{e}_t^T = (I + P.U_t)^{-1}.\underline{e}_t.\underline{e}_t^T.(I + U_t.P)$$

where $U_t = Q_t - V_t$. Then ,

$$f(P) = const + (P - P_{t-1}).(S_{t-1} + V_{t-1} - V_{t-1}.Q_t^{-1}.V_{t-1} + (I + P.U_t)^{-1}.\underline{e}_t.\underline{e}_t^T.(I + U_t.P)^{-1}$$

or , $f(P) = const^* + P.S_t$, where

$$S_t = S_{t-1} + (I + P_{t-1}.U_t)^{-1}.[\underline{e}_t.\underline{e}_t^T + U_t.(I + U_t.P_{t-1})].(I + U_t.P_{t-1})^{-1}$$

or , equivalently ,

$$S_t = Q_t^{-1}.V_{t-1}.[\underline{e}_t.\underline{e}_t^T + U_t.V_{t-1}^{-1}.Q_t].V_{t-1}.Q_t^{-1}$$

Appendix 7.1b - The Robust Filter Equations

Consider the following (partially) non-normal version of the multivariate D.L.M. model , given by ,

$$\underline{y}_t = F_t.\underline{\theta}_t + \underline{v}_t$$

$$\underline{\theta}_t = G.\underline{\theta}_{t-1} + \underline{w}_t \qquad \text{where :}$$

i) the observational noise $\underline{v}_t$ has an unimodal, symmetric and twice differentiable density.

ii) the system noise $\underline{w}_t$ has density approximately normal (or its variance W is small in comparison to the observational variance noise V) or equivalently , $\underline{\theta}_t / D_{t-1} \simeq N(\underline{a}_t, R_t)$.

Then , $(\underline{\theta}_t/D_t) \simeq N(\underline{m}_t, C_t)$, where the posterior moments for $\underline{\theta}_t$ are given by the following Robust Filter Equations :

$$\underline{m}_t = \underline{a}_t + B_t.g(\underline{y}_t)$$

$$C_t = R_t - B_t.H(\underline{y}_t).B_t^T$$

where the (vector) influence function g and the information matrix H are defined respectively by ,

$$g(\underline{y}_t) = -\frac{\partial}{\partial \underline{y}_t} ln\, p(\underline{y}_t/D_{t-1})$$

$$H(\underline{y}_t) = \frac{\partial}{\partial \underline{y}_t^T} g(\underline{y}_t)$$

and the covariance matrix B_t is given by

$$B_t = Cov(\underline{\theta}_t, \underline{y}_t/D_{t-1}) = R_t.F^T$$

.

Proof:

By Bayes Theorem,

$$\underline{m}_t = \int \underline{\theta}_t.p(\underline{\theta}_t/D_t)\, d\underline{\theta}_t = [p(\underline{y}_t/D_{t-1})]^{-1}.\int \underline{\theta}_t.p(\underline{y}_t/\underline{\theta}_t).p(\underline{\theta}_t/D_{t-1})\, d\underline{\theta}_t$$

and consequently ,

$$p(\underline{y}_t/D_{t-1}).(\underline{m}_t - \underline{a}_t) = \int p(\underline{y}_t/\underline{\theta}_t).(\underline{\theta}_t - \underline{a}_t).p(\underline{\theta}_t/D_{t-1})\, d\underline{\theta}_t$$

by an approximate normality assumption for $p(\underline{\theta}_t/D_{t-1})$,

$$\frac{\partial p(\underline{\theta}_t/D_{t-1})}{\partial \underline{\theta}_t} = -R_t^{-1}.(\underline{\theta}_t - \underline{a}_t).p(\underline{\theta}_t/D_{t-1})$$

and then ,

$$p(\underline{y}_t/D_{t-1}).(\underline{m}_t - \underline{a}_t) = -R_t.\int p(\underline{y}_t/\underline{\theta}_t).\frac{\partial p(\underline{\theta}_t/D_{t-1})}{\partial \underline{\theta}_t}\, d\underline{\theta}_t$$

Now , integrating by parts , we get

$$p(\underline{y}_t/D_{t-1}).(\underline{m}_t - \underline{a}_t) = -R_t.[0 - \int p(\underline{\theta}_t/D_{t-1}).\frac{\partial}{\partial \underline{\theta}_t} p(\underline{y}_t/\underline{\theta}_t)\, d\underline{\theta}_t]$$

and the r.h.s. became $R_t.\frac{\partial}{\partial \underline{\theta}_t} p(\underline{y}_t/D_{t-1}) = R_t.F_t^T.\frac{\partial}{\partial \underline{y}_t} p(\underline{y}_t/D_{t-1})$ or ,

$$\underline{m}_t - \underline{a}_t = R_t.F_t^T.\frac{\partial}{\partial \underline{y}_t} lnp(\underline{y}_t/D_{t-1})$$

, and the first filter equation is proved.

Also, in a similar manner for C_t , using integration by parts twice, we get $C_t = R_t - B_t.H(\underline{y}_t).B_t^T$.

Basic Properties:

i) In the normal case, the Robust Filter equations coincide with the standard bayesian analysis of the multivariate DLM .

Proof : Since , by hypothesis , $p(\underline{y}_t/D_{t-1}) \sim N(\underline{f}_t, Q_t)$, we have ,

$$ln\, p(\underline{y}_t/D_{t-1}) \propto -\frac{1}{2}.ln\,|Q_t| - \frac{1}{2}.(\underline{y}_t - \underline{f}_t)'.Q_t^{-1}.(\underline{y}_t - \underline{f}_t)$$

Then , by definition , we have

$$g(\underline{y}_t) = -\frac{\partial}{\partial \underline{y}_t} ln\, p(\underline{y}_t/D_{t-1}) = Q_t^{-1}.(\underline{y}_t - \underline{f}_t)$$

$$H(\underline{y}_t) = \frac{\partial}{\partial \underline{y}_t} g(\underline{y}_t) = Q_t^{-1}$$

Now , by the filter equations , we have

$$\underline{m}_t = \underline{a}_t + B_t.Q_t^{-1}.(\underline{y}_t - \underline{f}_t)$$

$$C_t = R_t - B_t.Q_t^{-1}.B_t^T = R_t - B_t.Q_t^{-1}.Q_t.Q_t^{-1}.B_t^T$$

Or , equivalently , since $A_t = R_t.F_t^T.Q_t^{-1} = B_t.Q_t^{-1}$ we have

$$\underline{m}_t = \underline{a}_t + A_t.(\underline{y}_t - \underline{f}_t)$$

$$C_t = R_t - A_t.Q_t.A_t^T$$

which are the standard updating equations for the DLM .

ii) Interpretation of g(.) and H(.)

By Bayes theorem , we have

$$\frac{\partial}{\partial \underline{y}_t} ln\, p(\underline{\theta}_t/\underline{y}_t) = \frac{\partial}{\partial \underline{y}_t} ln\, p(\underline{y}_t/\underline{\theta}_t) - E_{\underline{\theta}_t}\left\{\frac{\partial}{\partial \underline{y}_t} ln\, p(\underline{y}_t/\underline{\theta}_t)\right\}$$

We see that the influence of $\underline{y}_t$ on the posterior distribution will tend to dominate the combined influence of the previous observations when the value of $\frac{\partial}{\partial \underline{y}_t} ln\, p(\underline{y}_t/\underline{\theta}_t)$ is large relative to its expected value in the context of all previous observations . The negative of this derivative is called the influence function of the likelihood with respect to $\underline{y}_t$ [Ramsay & Novick(1980)] , the minus sign ensuring that the influence of the large observation will be positive .

In the normal case , as we can see from i) , the influence function $g(\underline{y}_t) = Q_t^{-1}.(\underline{y}_t - \underline{f}_t)$ is linear on the present data $\underline{y}_t$ and its predictive mean $\underline{f}_t$. As a consequence , the influence

function is unbounded . In a non-normal context as for instance in the Student t case , the influence function is bounded and its non-linearity operating on the residuals $\underline{e}_t = \underline{y}_t - \underline{f}_t$ deemphasizes the influence from large residuals . In particular , since $p(\underline{y}_t/D_{t-1})$ is unimodal , we see from the updating for $\underline{m}_t$ that the correction term $B_t.g(.)$ disappears (and the influence function will be zero) when the observation on $\underline{y}_t$ coincides with the mode of $p(\underline{y}_t/D_{t-1})$ and not the mean as in the normal filter .

Also , as $H(\underline{y}_t)$ by definition is a second derivative of a log- likelihood , it can be interpreted as a kind of Information matrix . In fact , the information about the random variable $\underline{y}_t$ contained in H is transfered to $\underline{\theta}_t$ through the covariance matrix B_t . This information matrix , as we have seen before , coincide with the predictive precision matrix Q_t^{-1} in the normal case and will be proportional to Q_t^{-1} in the Student t case .

iii) Other Properties :

Another way to look or to interpret the Robust filter is to see it as an approximation to the exact minimum variance filter where the state estimate is formed as a linear prediction corrected by a non-linear function of past and present observations . The approximation to the minimum variance filter becomes better and better as the ratio between the state and observational variances go to zero , when the relations will hold exactly . In particular , the system variance W will approach zero in this case and this corresponds to the use of a discounting factor close to 1 .

Further details or other properties about the Robust filter can be found for instance , in Masreliez(1975) or West(1981) .

Equivalently, the filter equations can be written as [West(1981)]

$$\underline{m}_t = \underline{a}_t + C_t.F_t^T.g(\underline{y}_t - F_t.\underline{a}_t)$$

$$C_t^{-1} = R_t^{-1} + F_t^T.H(\underline{y}_t - F_t.\underline{a}_t).F_t$$

where g(.) and H(.) are defined as before.

Appendix - 7.2 : An Approximated Likelihood Function

Now , consider (7.21b) with the log-density given by (7.21) , where the evaluations are easier if we consider each element of V individually. Typical forms to appear (see expression 7.21)

are the partials of the log of a determinant and of a quadratic form involving an inverse matrix ; such derivatives can be expressed as (see Press, pg 41 or Maybeck, pg 79)

$$\frac{\partial ln\,|A|}{\partial \nu_k} = \frac{1}{|A|}.\frac{\partial |A|}{\partial \nu_k} = tr\{A^{-1}.\frac{\partial A}{\partial \nu_k}\}$$
$$\frac{\partial A^{-1}}{\partial \nu_k} = -A^{-1}.\frac{\partial A}{\partial \nu_k}.A^{-1}$$

where ν_k is the k^{th} element of vec B , with A and B square matrices (in our case , B represents V and A represents the covariance matrices C_t or Q_j) .

Using the above two derivative formulaes , equation (7.21) becomes

$$\frac{\partial}{\partial \nu_k} ln\, p(\underline{\theta}_t, D_t/V) = -\frac{1}{2}.tr\{C_t^{-1}.\frac{\partial C_t}{\partial \nu_k}\} - \frac{1}{2}.\sum_{j=1}^{t} tr\,\{Q_t^{-1}.\frac{\partial Q_t}{\partial \nu_k}\}$$

$$+\frac{1}{2}.\sum_{j=1}^{t}(\underline{y}_j - F^T.\underline{a}_j)^T.Q_j^{-1}.\frac{\partial Q_j}{\partial \nu_k}.Q_j^{-1}.(\underline{y}_j - F^T.\underline{a}_j) + 2.\sum_{j=1}^{t}\frac{\partial \underline{a}_j}{\partial \nu_k}.F^T.Q_j^{-1}.(\underline{y}_j - F^T.\underline{a}_j) = 0$$

where we should notice that the two other terms in 7.21 involving $(\underline{\theta}_t - \underline{m}_t)$ are null because $\underline{\theta}_t$ is set equal to the ML estimate $\hat{\underline{\theta}}_t = \underline{m}_t$ (see equation 7.22) .

Also , using well known properties about the trace of a matrix (such as $a^T.b = tr(a.b^T) = tr(b.a^T)$ for any vectors a and b) , the likelihood equation above can be rewritten as :

$$tr\{C_t^{-1}.\frac{\partial C_t}{\partial \nu_k}\} - 2.\sum_{j=1}^{t}\frac{\partial \underline{a}_t^T}{\partial \nu_k}.F.Q_j^{-1}.\underline{u}_j + \sum_{j=1}^{t} tr\{[Q_j^{-1} - Q_j^{-1}.\underline{u}_j.\underline{u}_j^T.Q_j^{-1}].\frac{\partial Q_j}{\partial \nu_k}\} = 0$$

where $\underline{u}_j = \underline{y}_j - F.\underline{a}_j$ and ν_k is the k^{th} element of vec V .

Now , we should use the knowledge that V varies much slower than the quantities related with $\underline{\theta}_t$ in order to get equations more suitable for sequential estimation. This implies that one adequate model is such that V is essentially constant over any given interval of N sample periods ; that is , at a given time t , the parameters in V are supposed to have remained constant since the time t-N+1. This sort of fixed-length memory version of the ML estimator is obtained by rewriting the former equations with the lower limits in the sums equal to t-N+1 instead of 1 .

It is important to mention however that there is no general closed form solution to this last equation and some sort of iterative solution or direct approximation of the likelihood equation is necessary.

To enhance online applicability , not only can an iterative solution procedure be approximated (see for instance , Maybeck(1982) pp 80-120) but the likelihood equations themselves

can be approximated as well. As pointed out by Maybeck(1982) , analyses based on simulations or other methods can indicate the relative sensitivities of individual terms in the likelihood equation to parameter values , and the less sensitive terms can be neglected.

With these ideas in mind , we neglect the first two terms in the former equation , because it represents the sensitivity of state variable statistics (specifically $\underline{a}_t$ and C_t) on V , what we know is considerably robust , giving the following approximated likelihood equation :

$$\sum_{j=t-N+1}^{t} tr\{[Q_j^{-1} - Q_j^{-1}.\underline{u}_j.\underline{u}_j^T.Q_j^{-1}].\frac{\partial Q_j}{\partial \nu_k}\}\Bigg|_{\underline{\theta}_t=\underline{m}_t} = 0$$

which is exactly equation (7.23) of section 7.3.

Appendix - 7.3 : Linear Bayes Theory and related results

Some Basic Concepts :

Lets f(Y) be the Bayes Estimator for a certain random variable X . Then , the Bayes risk is defined by ,

$$r(f) = E\{f(Y) - X\}^2 = E_Y E_{X/Y}\{f(Y) - X\}^2 = E_Y d(f/Y)$$

where d(f/Y) is the posterior expected squared error , i.e., the posterior expected loss for a quadratic loss function . Clearly , r(f) is minimised completely by minimizing d(f/Y) for each Y , what gives f(Y) = E(X/Y) .

However , this is an arbitrarily complicated function of Y and requires full Bayesian analysis to derive . Linear estimation simplifies the derivation by restricting f(Y) to the linear class $f(Y) = a + b^T.Y$. Then , a and b are chosen to minimise r(f) , what gives the optimal estimate

$$\hat{f}_X(Y) = \hat{a} + \hat{b}^T.Y$$

where the coeficients $\hat{a}$ and $\hat{b}$ are given respectively by

$$\hat{a} = E(X) - Cov(X,Y).[Var(Y)]^{-1}.E(Y)$$
$$\hat{b} = Cov(X,Y).[Var(Y)]^{-1}$$

Also , $V_X = r(\hat{f})$ is a measure of how good $\hat{f}(Y)$ is . It is expected squared error , but notice that it is the unconditional expectation , i.e. , the prior expectation of posterior expected squared error .

Now , suppose X is a vector . In general , we could estimate each component X_i using a linear function of a different predictor vector Y_i given by $f_i(X_i) = a_i + b_i^T.Y_i$. Then , a_i and b_i will be chosen to minimize $r_i(f_i) = E\{f_i(Y) - X_i\}^2$ and a measure of accuracy is provided by $V_{X_i} = r_i(\hat{f}_i)$, what gives a kind of expected posterior variance for $\hat{f}_i$.

We would also like an analogue of covariance , and this is obtained as follows . Let ,

$$f(Y) = \begin{pmatrix} f_1(Y_1) \\ \vdots \\ f_k(Y_k) \end{pmatrix} \qquad \& \qquad \hat{f}(Y) = \begin{pmatrix} \hat{f}_1(Y_1) \\ \vdots \\ \hat{f}_k(Y_k) \end{pmatrix}$$

and define $\quad r(f) = E\{f(Y) - X\}\{f(Y) - X\}^T \quad$ and $\quad V_X = r(\hat{f})$.

The diagonal elements of V_X will be V_{X_i} and the off-diagonal elements will be the expected cross-products of errors , which will correspond to covariances .

Then , the (optimal) linear Bayes estimate $M = f(Y)$ of the random vector X (given the data vector Y) and its associated expected squared error C are given respectively by:

$$M = E(X) + A.(Y - E(Y))$$

$$C = Var(X) - A.Var(Y).A^T$$

where $\quad A = Cov(X,Y).[Var(Y)]^{-1}$.

Some second order moments and related results

The objective here is to develop an expression for $Var(\underline{y}_t \otimes \underline{y}_t / \underline{\pi}_t)$ where $(\underline{y}_t / \underline{\pi}_t) \sim (\underline{\psi}_t, \Sigma_t)$ and this is done as follows . After some basic results about Jacobian matrices we present a result (Lemma) for the desired variance in the case of $\underline{\psi}_t = \underline{0}$ and $\Sigma = I$ and then the general result .

Jacobian transformations : the Commutation and the Symmetrization matrices .

Definition : The Jacobian matrix of the transformation from the real mxn matrix A to its transpose A^T (or equivalently , ... permutation from vec A to vec A^T) ,

$$K_{mn} = \frac{\partial\, vec\, A^T}{\partial\, (vec\, A)^T} \qquad (\text{ or , equivalently , } K_{mn}.vec\, A = vec\, A^T)$$

is called a Commutation Matrix .

This matrix can be represented as ,

$$K_{mn} = \sum_{i=1}^{m}\sum_{j=1}^{n}(H_{ij} \otimes H_{ij}^T)$$

where H_{ij} is a mxn matrix with 1 in its ij^{th} position and zeros elsewhere .

Commutatibility : The key property of the commutation matrix (and the one from which it derives its name) enables us to interchange or commutate the two matrices of a Kronecker product . Let A be a mxn matrix and B a pxq matrix ; then

$$K_{pm}.(A \otimes B) = (B \otimes A).K_{qn}$$

In particular , if b is a px1 vector , $K_{pm}.(A \otimes b) = b \otimes A$.

Symmetrization : Closely related to the commutation matrix is the Symmetrization Matrix N_n . It transforms an arbitrary nxn matrix A into the symmetric matrix $\frac{1}{2}.(A + A^T)$. Naturaly , they are related by the expression ,

$$N_n = \frac{1}{2}.(I_{n^2} + K_{nn})$$

The explicit form of N_n is derived from K_{nn} . For instance , for $n = 2$, we have

$$L_2 = 2.N_2 = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

Lemma : $Var(\underline{u} \otimes \underline{u}) = 2.N_n$ where $\underline{u} \sim (0, I_n)$ and N_n is the symmetrization matrix.

Proof : Without loss of generality , suppose A is a nxn symmetric matrix . Then , considering as an approximation that $\underline{u}$ has the same moments as the normal distribution up to 4^{th} order , and using well known results for variance of a quadratic form (see , for instance , Seber,G.A.F. pg 41) , we have :

$$Var(\underline{u}'.A.\underline{u}) = 2.tr\,A^2 = tr\,A'.A + tr\,A^2$$

Now , using the property that $tr\,A'.B = (vec\,A)'.vec\,B$ and the definition of the commutation matrix , we have ,

$$Var(\underline{u}'.A.\underline{u}) = (vec\,A)'.vec\,A + (vec\,A')'.vec\,A$$

$$Var(\underline{u}'.A.\underline{u}) = (vec\,A)'.(I_{n^2} + K_{nn}).vec\,A \qquad (*)$$

On the other hand , $\underline{u}'.A.\underline{u} = vec\,\underline{u}'.A.\underline{u}$, and using the property that $vec\,A.X.B = (B' \otimes A).vec\,X$ we have

$$Var(\underline{u}'.A.\underline{u}) = (vec\,A)'.Var(\underline{u} \otimes \underline{u}).vec\,A \qquad (**)$$

Comparing $(*)$ and $(**)$, we have , finally

$$Var\,(\underline{u} \otimes \underline{u}) = I_{n^2} + K_{nn} = 2.N_n$$

Now , let's extend this result to the case where we have $(\underline{y}_t \,/\, \underline{\pi}_t) \sim (\underline{\psi}_t\,, \Sigma_t)$ and we want $Var\,(\underline{y}_t \otimes \underline{y}_t \,/\, \underline{\pi}_t)$. This is done as follows:

Conditionally on $\underline{\pi}_t$, we have $\underline{y}_t = \Sigma_t^{\frac{1}{2}}.\underline{u}_t + \underline{\psi}_t$, where $\underline{u}_t \sim (\underline{0}, I_p)$, and then

$$\underline{y}_t \otimes \underline{y}_t = \Sigma_t^{\frac{1}{2}}.\underline{u}_t \otimes \Sigma_t^{\frac{1}{2}}.\underline{u}_t + \Sigma_t^{\frac{1}{2}}.\underline{u}_t \otimes \underline{\psi}_t + \underline{\psi}_t \otimes \Sigma_t^{\frac{1}{2}}.\underline{u}_t + \underline{\psi}_t \otimes \underline{\psi}_t$$

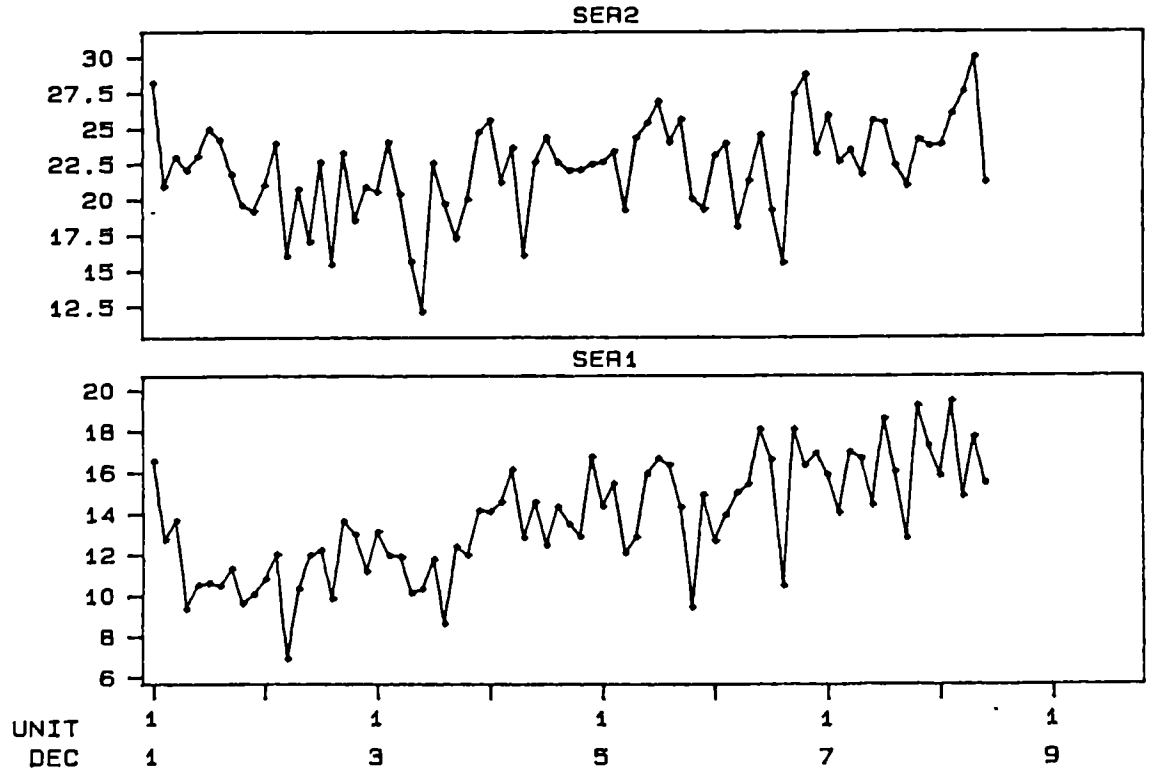
Using twice the property that $(A \otimes B).(C \otimes D) = A.C \otimes B.D$ and also the commutability and symmetrization properties , we have

$$\underline{y}_t \otimes \underline{y}_t = (\Sigma_t^{\frac{1}{2}} \otimes \Sigma_t^{\frac{1}{2}}).(\underline{u}_t \otimes \underline{u}_t) + 2.N_p.(\Sigma_t^{\frac{1}{2}} \otimes \underline{\psi}_t).\underline{u}_t + \underline{\psi}_t \otimes \underline{\psi}_t$$

Using the Lemma's result and the matrix product property twice again , we have finally ,

$$Var\,(\underline{y}_t \otimes \underline{y}_t \,/\, \underline{\pi}_t) = L_p.[\Sigma_t \otimes \Sigma_t + (\Sigma_t \otimes \underline{\psi}_t.\underline{\psi}_t').L_p] \qquad \text{where } L_p = 2.N_p$$

FIG. 7.1 TWO DIMENSIONAL SIMULATED TIME SERIES



TAB 7.1 – BIVARIATE SIMULATED TIME SERIES

SER1

16.563 12.712 13.674 9.3503 10.507 10.596 10.445 11.289 9.6236 10.061 10.78
12.015 6.9074 10.33 11.969 12.215 9.8493 13.648 12.983 11.159 13.112
11.938 11.883 10.12 10.312 11.772 8.632 12.382 11.962 14.156 14.086
14.569 16.135 12.815 14.561 12.446 14.332 13.477 12.847 16.757 14.321
15.461 12.07 12.832 15.954 16.655 16.33 14.31 9.4395 14.892 12.658
13.928 15.018 15.435 18.097 16.576 10.454 18.088 16.305 16.922 15.883
14.042 16.971 16.654 14.414 18.579 15.993 12.77 19.222 17.27 15.799
19.425 14.819 17.697 15.449

SER2

28.209 20.879 22.949 22.018 22.992 24.893 24.137 21.705 19.544 19.096 21.001
23.916 15.997 20.717 16.996 22.579 15.407 23.229 18.482 20.838 20.484
23.99 20.338 15.606 12.114 22.536 19.666 17.237 19.988 24.717 25.534
21.157 23.618 16.05 22.61 24.357 22.577 21.989 22.053 22.467 22.613
23.347 19.178 24.328 25.345 26.894 23.981 25.608 19.981 19.233 23.055
23.927 18.047 21.303 24.505 19.182 15.479 27.43 28.789 23.21 25.895
22.598 23.444 21.684 25.546 25.371 22.311 20.92 24.153 23.717 23.803
26.045 27.609 30.012 21.17

FIG. 7.2 ONE STEP AHEAD FORECASTING ERROR SERIES 1

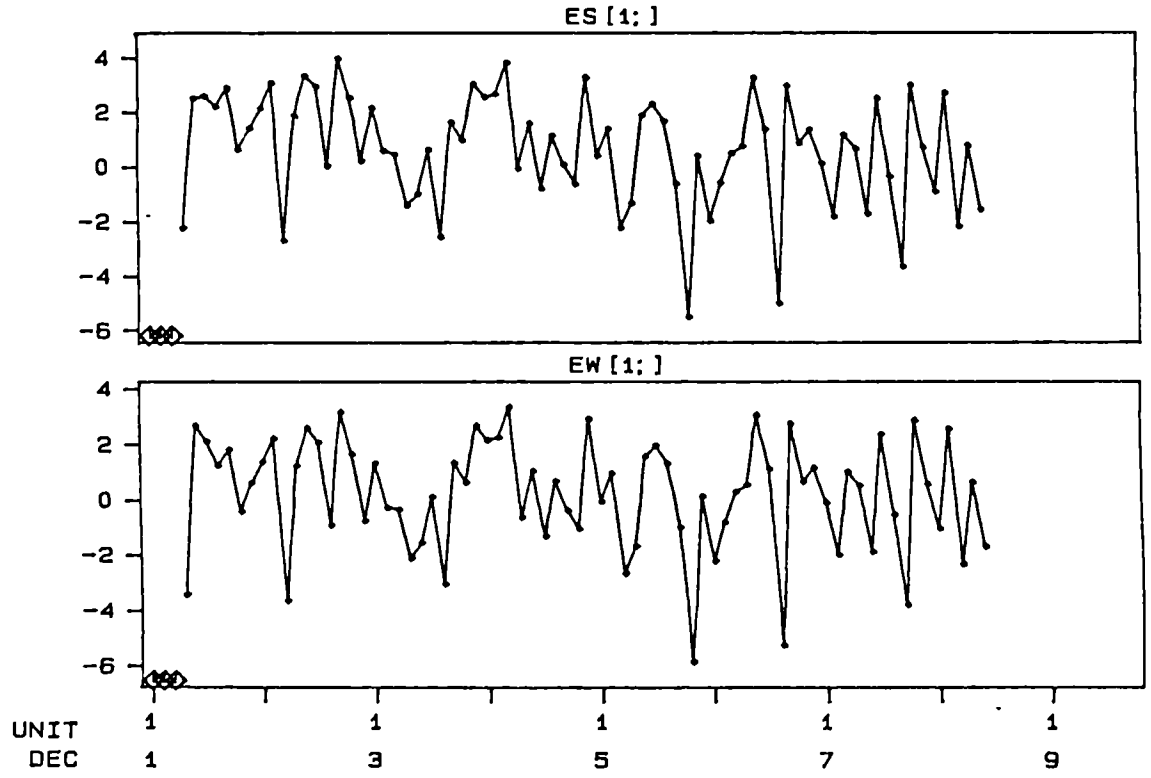
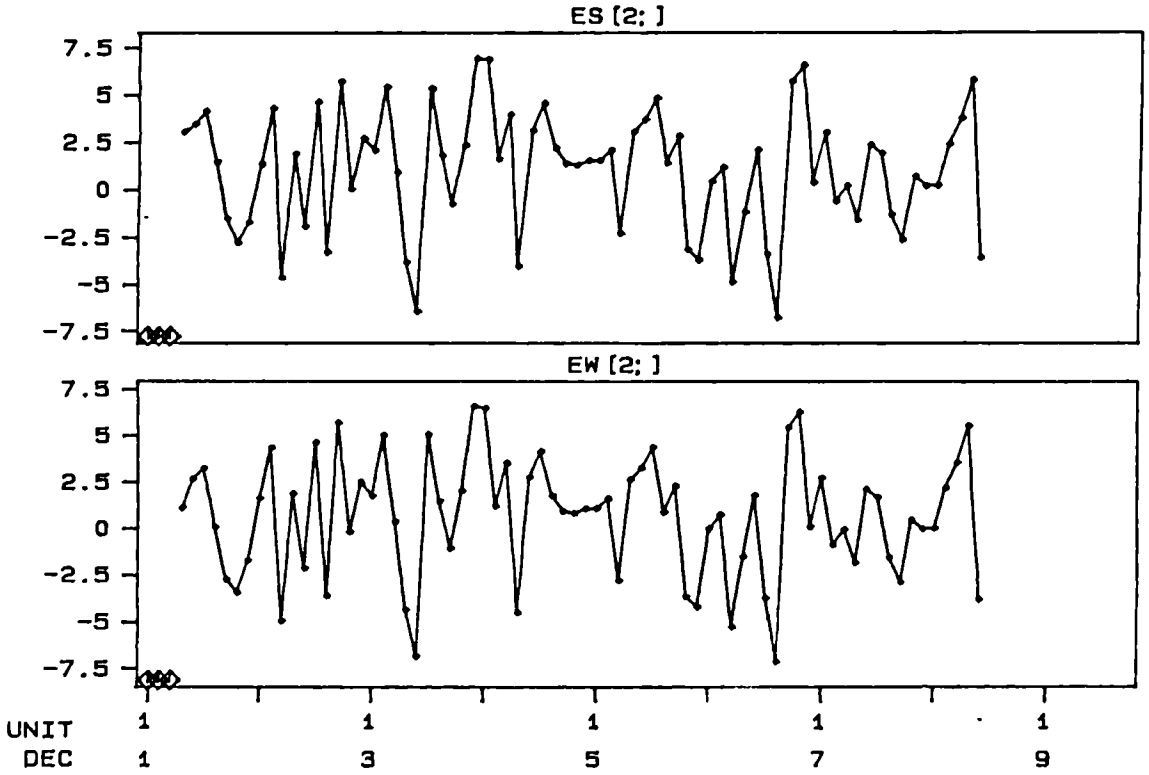


FIG. 7.3 ONE STEP AHEAD FORECASTING ERRORS SERIES 2



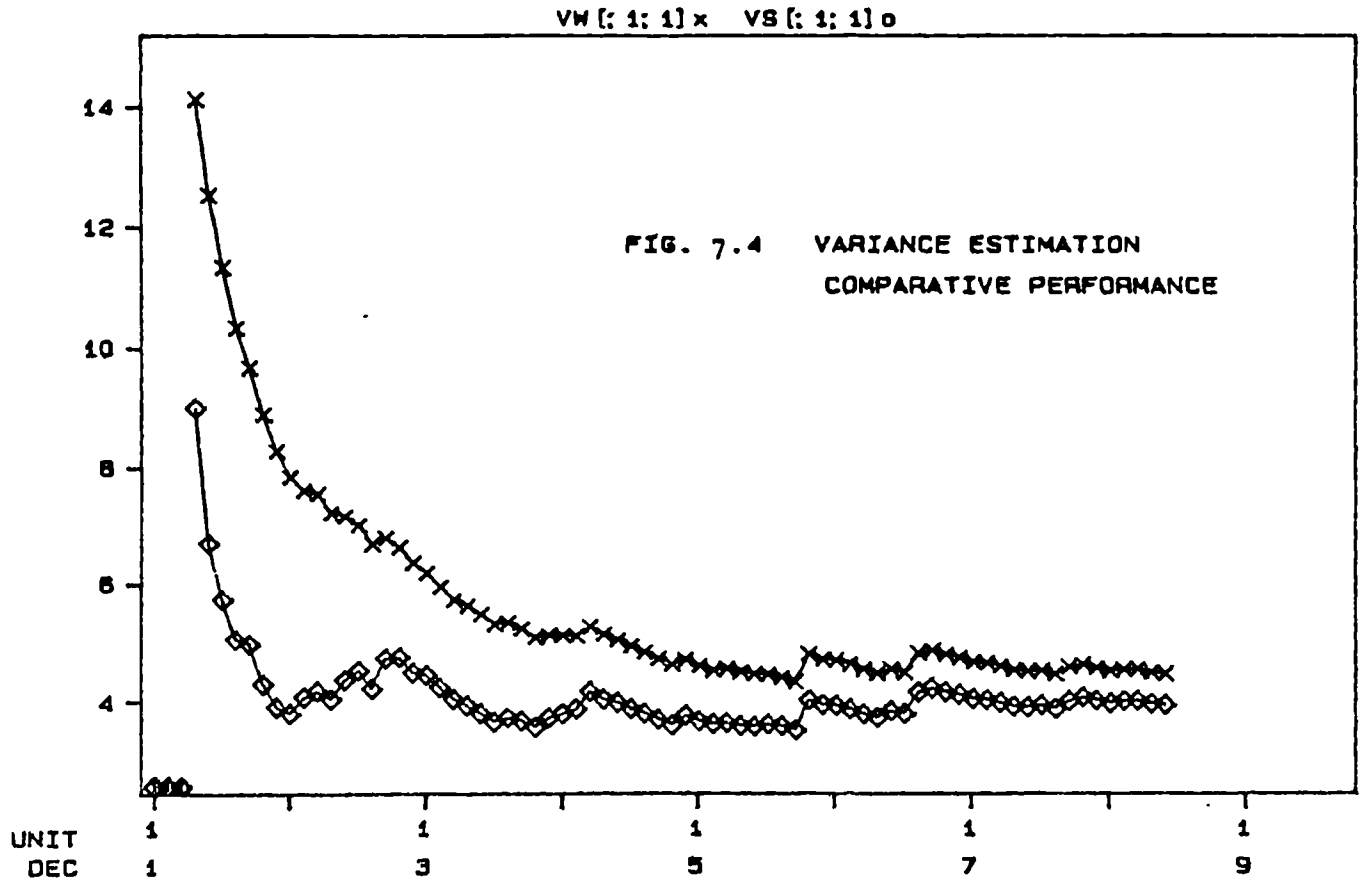
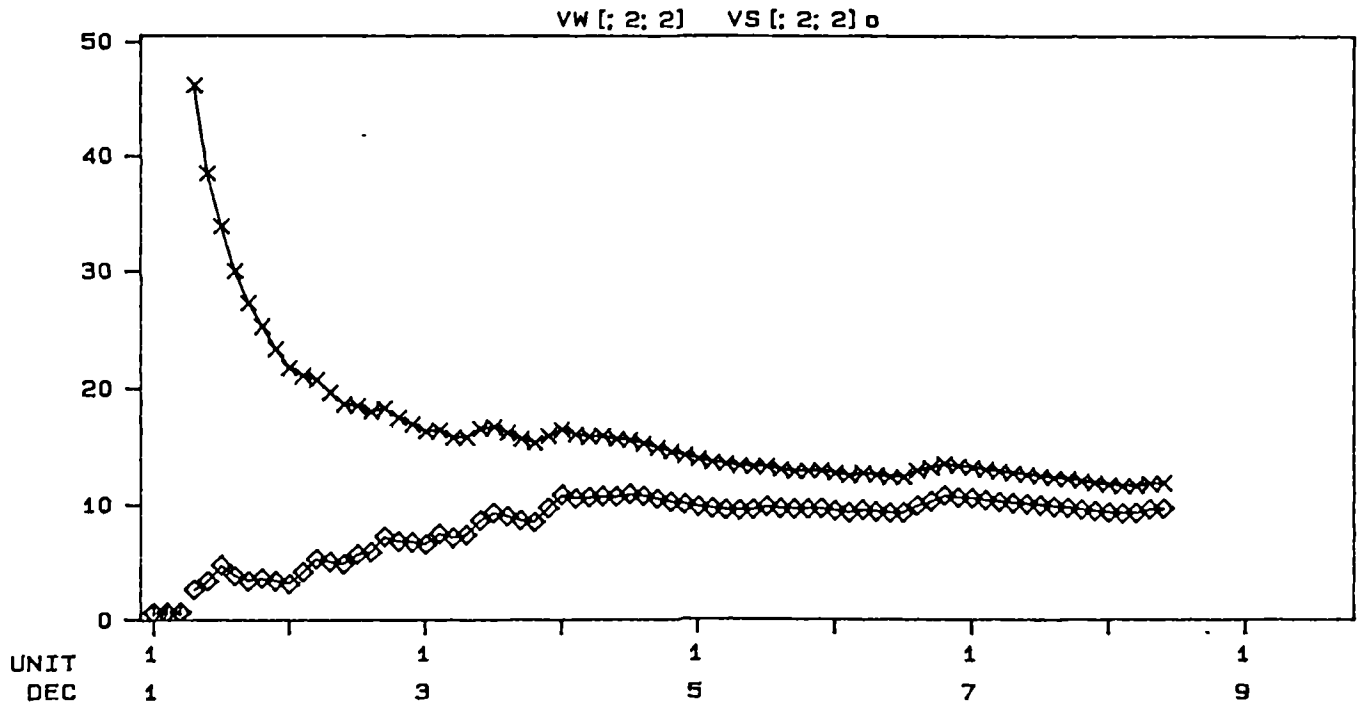


FIG. 7.4 VARIANCE ESTIMATION
COMPARATIVE PERFORMANCE

FIG. 7.5 VARIANCE ESTIMATION COMPARATIVE PERFORMANCE SERIES 2



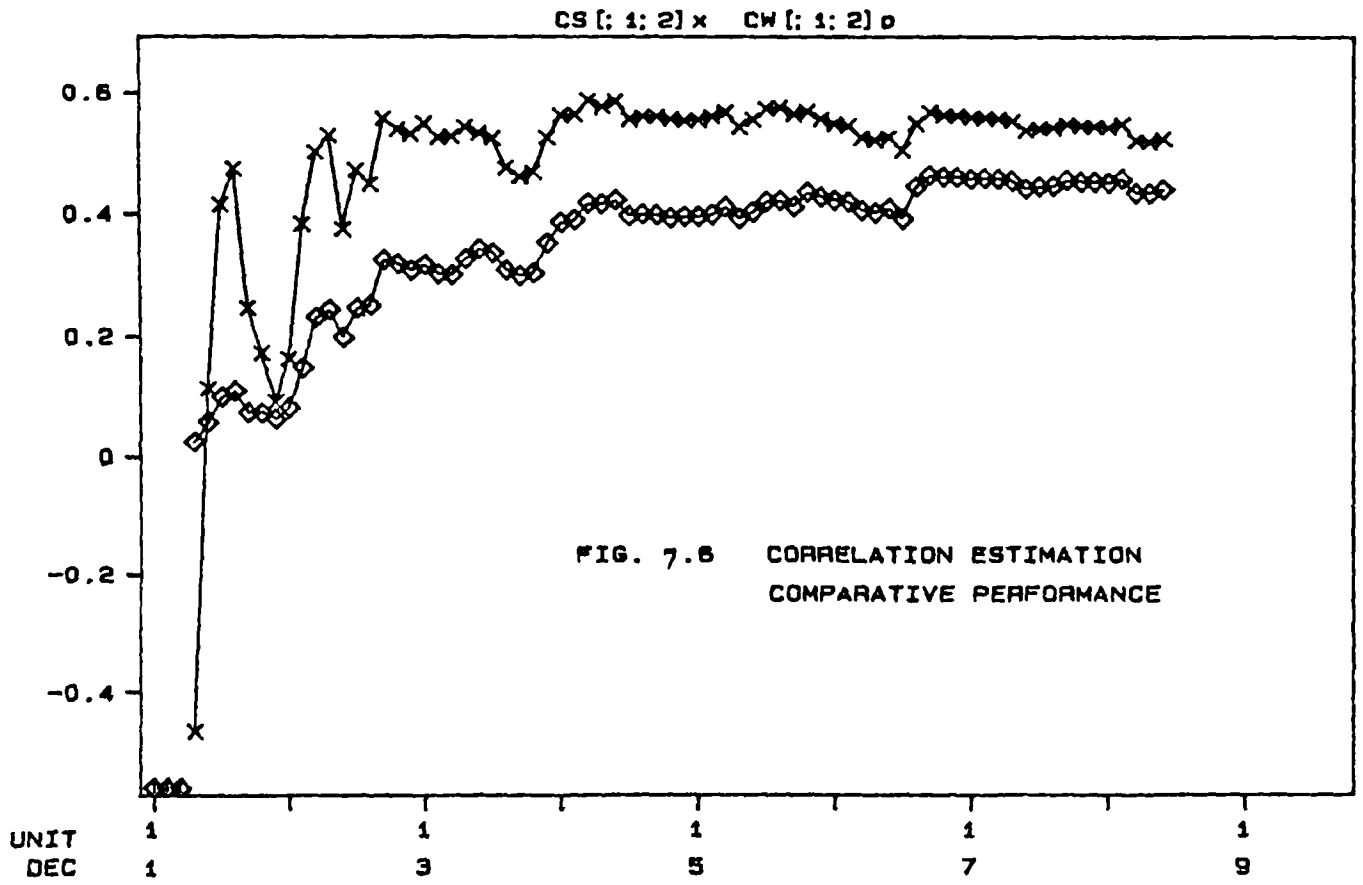


FIG. 7.6 CORRELATION ESTIMATION
COMPARATIVE PERFORMANCE

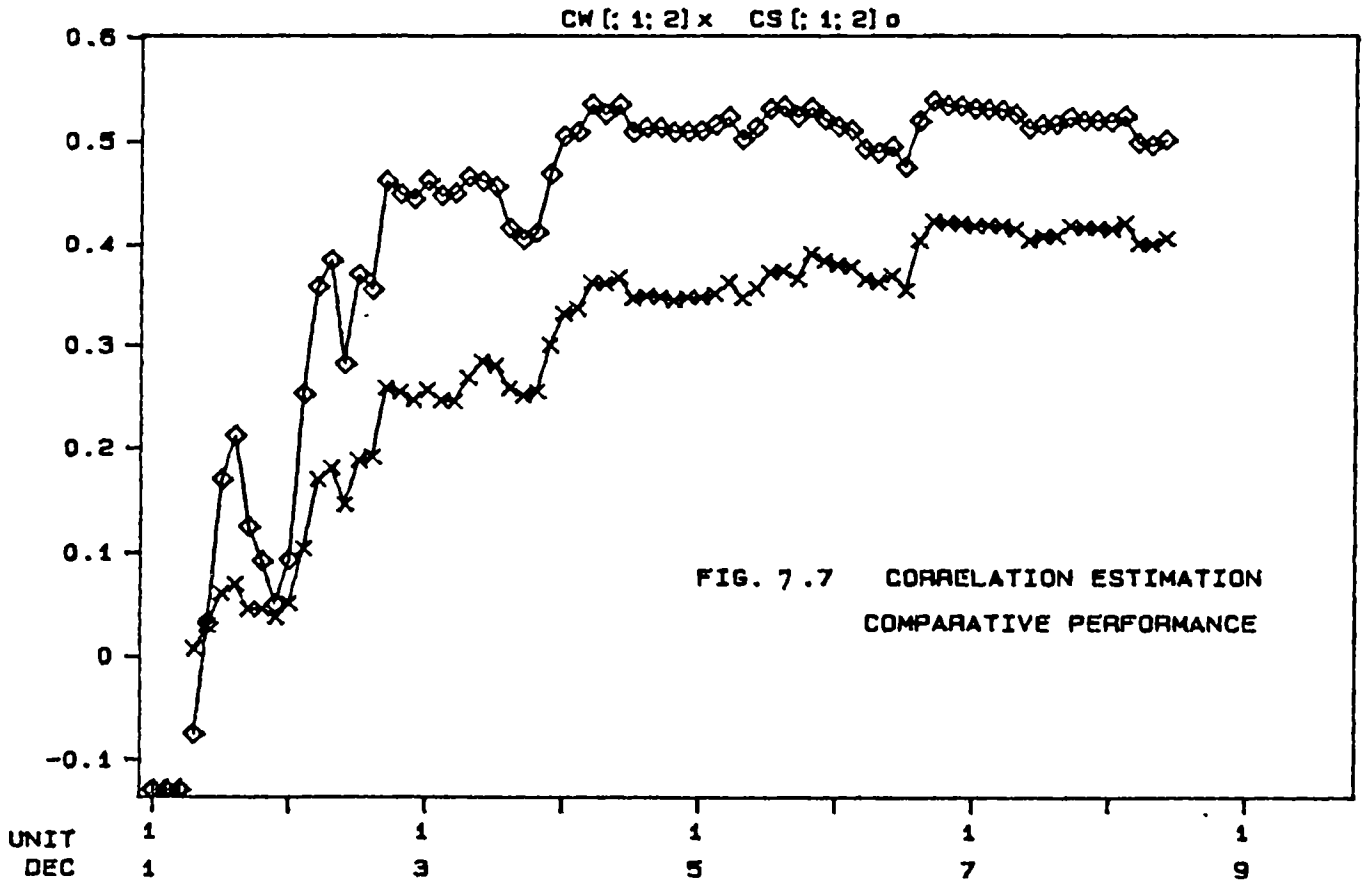


FIG. 7.7 CORRELATION ESTIMATION
COMPARATIVE PERFORMANCE

FIG. 7.8 CONVERGENCE TO THE STANDARD CONJUGATE ANALYSIS ALGORITHM

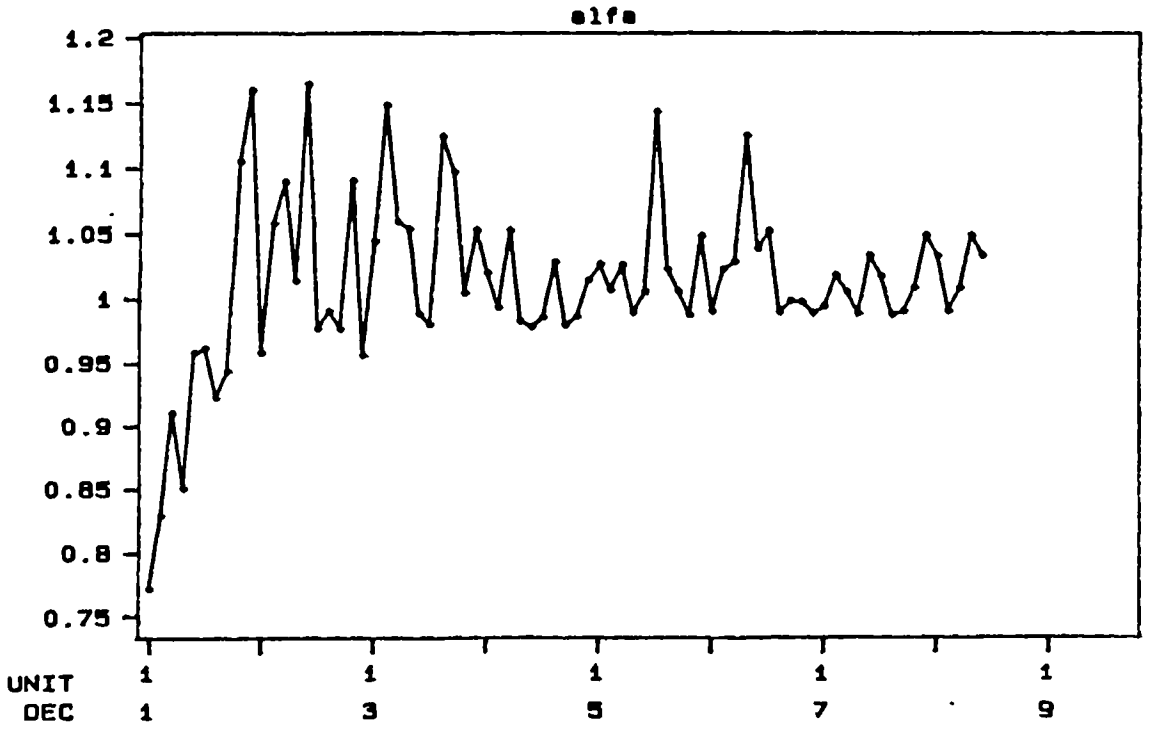
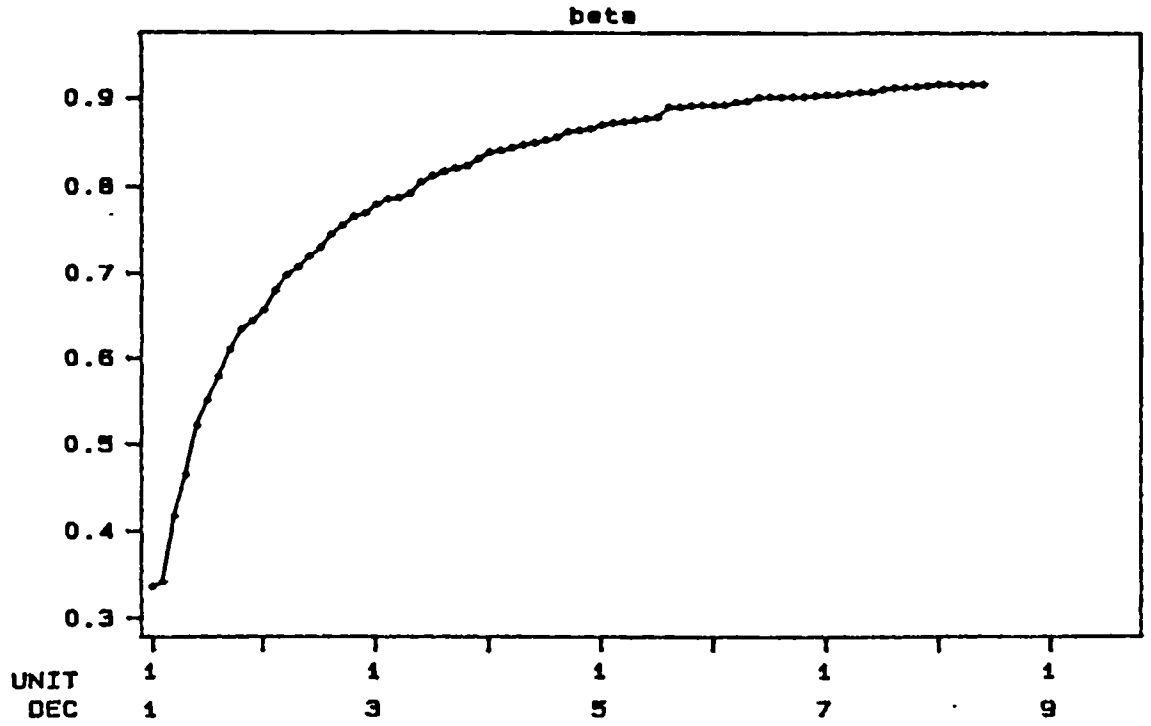


FIG. 7.9 MONITORING THE BETA COEFFICIENT EVOLUTION IN TIME



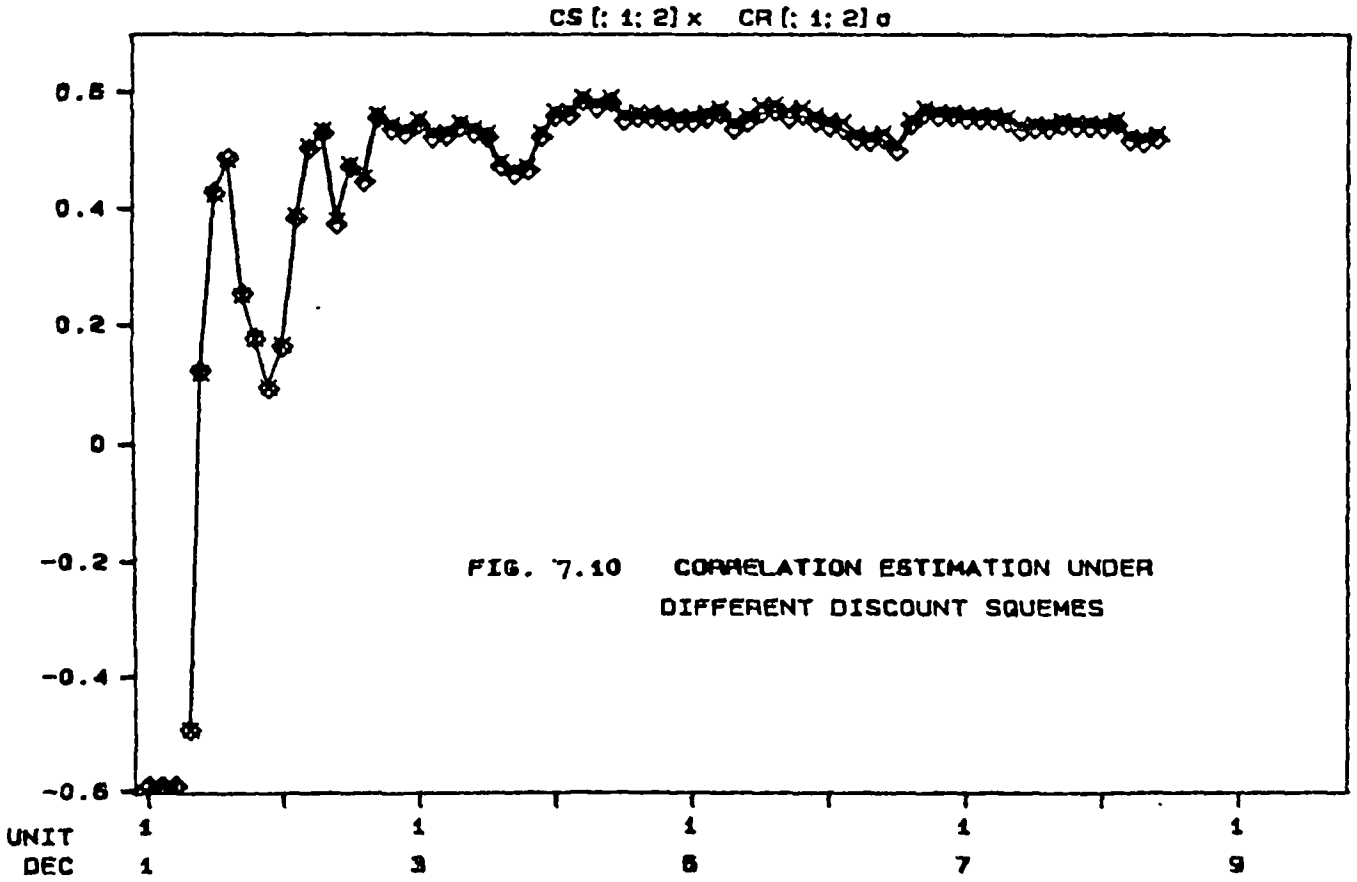


FIG. 7.10 CORRELATION ESTIMATION UNDER DIFFERENT DISCOUNT SQUEMES

CHAPTER 8

SOME SPECIAL MODELLING ASPECTS

8.1 - Introduction .

In this chapter we explore some important modelling aspects in conection with the multivariate DLM model , which is a very general and flexible modelling framework , in order to extend even more the range of possible applications for such models .

We should not forget however that the methods and models covered in the last three chapters of this thesis , were founded upon the assumptions of linearity and normality , and this can be unjustifiable in some special applications .

As discussed previously in chapter 3 of this thesis for univariate models , the possibility of extending the DLM structure in order to permit the introduction of non-linearities in the observation or system equation can be of extreme importance , as for instance , in the case of the seasonal growth multiplicative model and others .

With these ideas in mind , in the next section of this chapter , a non-linear extension of the methodology developed in chapter 6 for multivariate DLM analysis is introduced , as well as the specific case of the multivariate Seasonal Growth Multiplicative - SGM model . In fact , this new multivariate model extends not only the methods of chapter 6 , but also it extends the scalar SGM model of chapter 3 to the vectorial case .

Another kind of extension of multivariate DLM's useful in some practical applications is related with the modelling of some non-normal data , as for instance , time series data consisting of proportions or compositions of a whole (compositional time series) . This is possible through the use of some special transformations such as the logistic log-ratio transformation - Aitchison(1986) , as presented in section 8.3 of this chapter .

Finally , in section 8.4 of this chapter , one of the most popular methods in the econometric literature about multivariate time series modelling and forecasting , the so called BVAR (Bayesian Vector Auto-Regressive) model - Litterman(1980) , is analysed as a special case of common components multivariate DLM . Also , a possible way of extending this model to a more general framework is discussed .

8.2 - Modelling Non-Linearities

We present here an extension of the model formulation of chapter 6 in order to permit the introduction of non-linearities in the observation or system equation, making the range of possible applications for such models even wider. In the first couple of sub-sections we introduce a new class of non-linear multivariate models with the corresponding analysis and updating equations. In the following sub-section , the important case of the multivariate Seasonal Growth Multiplicative model is considered in detail .

8.2.1 - General formulation and some implementation aspects .

Concretely , we consider here a more general (non-linear) relation of link between the process parameter $\underline{\theta}_t$ and the mean response parameter $\underline{\mu}_t$ in order to extend the multivariate D.L.M. model of chapter 6 as follows :

Definition : A general multivariate (normal) Dynamic Non-Linear Model - D.N.L.M. for a vector of observations $\underline{y}_t$ of dimension d made at intervals at times t = 1,2,.. is defined by the following equations :

$$\text{i) observation equation :} \quad \underline{y}_t = g(\underline{\theta}_t) + \underline{v}_t \quad , \quad \underline{v}_t \sim N(\underline{0}, V_t) \tag{8.1}$$

$$\text{ii) system equation :} \quad \underline{\theta}_t = G.\underline{\theta}_{t-1} + \underline{w}_t \quad , \quad \underline{w}_t \sim N(\underline{0}, W_t) \tag{8.2}$$

where :

i) g is a (non-linear) differentiable function.

ii) all the other elements are defined as before.

It is clear that the formulation of equations (5.1)-(5.2) is a particular case of (8.1)-(8.2) in the special situation when g is a linear function.

Now, in order to make such model formulation operational , we introduce more structure in the above model definition as follows . First , we consider briefly a direct way of implementing the dynamic non-linear model given by the equations (8.1)-(8.2) based on an extension of the methodology of section 6.2 as follows .

Definition : A scaled version of the multivariate (normal) D.N.L.M. of this section is defined

by the equations (8.1)-(8.2) plus the following distributional assumptions :

$$(i) \quad \text{likelihood function :} \quad (\underline{y}_t / \underline{\mu}_t , V = S^2) \sim N(\underline{\mu}_t ; S^2) \tag{8.3}$$

$$(ii) \quad \text{prior distribution :} \quad (\underline{\mu}_t / D_{t-1} , V) \sim N(\underline{f}_t ; R_t^*) \tag{8.4}$$

$$(V^{-1} / D_{t-1}) \sim W(U_{t-1}^{-2} ; n_{t-1}) \tag{8.4a}$$

where :

(i') $\underline{\mu}_t = g(\underline{\theta}_t)$ where g is a differentiable function , $\underline{\mu}_t$ is the mean response parameter and $\underline{\theta}_t$ is the process parameter as usual .

(ii') Given V and D_{t-1} , $\underline{f}_t$ is the prior mean and $R_t^* = S.\Sigma_t.S$ is the prior variance matrix for $\underline{\mu}_t$, where $\Sigma_t = Var\{\underline{\mu}_t / D_{t-1} , V = I\}$. Also , U_{t-1}^{-2} is the prior observational precision matrix and n_{t-1} is the corresponding prior degree of freedom (d.f.) .

As the basic model structure is defined in terms of the mean response parameter , the model analysis for the scaled DNLM is structuraly analogous to the one developed in section 6.2.1 . The updating equations however , are the same as in section 6.2 only for the steps 1 and 3. In the steps 2 and 4 we redefine the elements $\underline{f}_t$, Σ_t and A_t^1 as follows :

$$\underline{f}_t = E[g(\underline{\theta}_t)/D_{t-1}, V] \tag{8.5}$$

$$\Sigma_t = Var[g(\underline{\theta}_t)/D_{t-1}, V = I] \tag{8.6}$$

$$A_t^1 = Cov[\underline{\theta}_t , g(\underline{\theta}_t)] . \Sigma_t^{-1} \tag{8.7}$$

These equations are implemented in practice , without major difficulties (as exemplified in section 8.2.3 for the SGM model) , and are functions of $\underline{a}_t$, R_t and g. The full updating equations for this model are presented in the Appendix 8.1 of this chapter [equations (8.8)-(8.25)] .

8.2.2 - An Alternative Implementation .

Although the procedures described in the last sub-section of this chapter for the implementation of the multivariate DNLM are perfectly valid and operational , they suffer from the same limitations as the methods of section 6.2 , where the posterior distribution for $\underline{\theta}_t$ is available only conditionally on a given V . For these reasons , we develop here an alternative way of implementing the model given by the equations (8.1)-(8.2) based on an extension of the methodology presented in section 6.3 to the non-linear case .

Before introducing the main model framework , we commence by presenting some introductory Definitions :

(i) The equivalent to the F element in a multivariate *DLM* context is *defined for* the more general framework of equations (8.1)-(8.2) as follows :

$$\hat{F}_t = \left. \frac{\partial g(\underline{\theta}_t)}{\partial \underline{\theta}_t} \right|_{\underline{\theta}_t = \underline{a}_t} \tag{8.26}$$

where $\underline{a}_t$ is the most recent estimate of $\underline{\theta}_t$. As expected , in the special case of linear models , where $g(\underline{\theta}_t) = F_t.\underline{\theta}_t$, we have that $\hat{F}_t = F_t$.

(ii) For the model structure of equations (8.1)-(8.2) , the generalized observability matrix $\hat{T}_t$ of dimension kdxp is defined by

$$\hat{T}_t = \begin{pmatrix} \hat{T}_{1t} \\ \hat{T}_{2t} \\ \vdots \\ \hat{T}_{kt} \end{pmatrix} = \begin{pmatrix} \hat{F}_t \\ \hat{F}_t.G \\ \vdots \\ \hat{F}_t.G^{k-1} \end{pmatrix} \tag{8.27}$$

where $\hat{F}_t$ is given by (i) and k is the maximum parametric dimension of the marginal models (d is the dimension of the observations and p is the dimension of the process parameters) . In the case of linear models , since $\hat{F}_t = F$, we have that $\hat{T}_t = T$, where T is the observability matrix given by the equation (6.21) .

Also , as a consequence , the corresponding to the matrix S_T given by equation (6.27) will be the matrix $\hat{S}_T$ of dimension kdxp defined by

$$\hat{S}_T = \begin{pmatrix} S^*.\hat{T}_{1t} \\ \vdots \\ S^*.\hat{T}_{kt} \end{pmatrix} \tag{8.28}$$

where S^* is such that $V = S^*.V_{t-1}.S^*$.

Definition : A scaled version of the multivariate (normal) DNLM of this section for a vector of observations $\underline{y}_t$ of dimension d made at times t=1,2,.. is defined by the equations (8.1)-(8.2) with the following distributional structure :

i) Prior distribution for V : $(V^{-1} / D_{t-1}) \sim W(d_{t-1}; n_{t-1})$ where d_{t-1} & n_{t-1} are respectively the shape parameter and the d.f. in the Wishart distribution such that $V_{t-1} = d_{t-1}/n_{t-1}$.

ii) Joint prior distribution for $\underline{\theta}_t$ and $\underline{y}_t$:

$$\begin{pmatrix} \underline{\theta}_t \\ \underline{y}_t \end{pmatrix} \Big| \; D_{t-1}, V = S^*.V_{t-1}.S^* \sim N\left\{ \begin{pmatrix} \underline{a}_t \\ \underline{f}_t \end{pmatrix}; \begin{pmatrix} \hat{R}_T & \hat{R}_T.\hat{F} \\ . & \hat{F}.\hat{R}_T.\hat{F} + V_{t-1} \end{pmatrix} \right\} \tag{8.29}$$

where $\hat{R}_T = \hat{T}_t^{-}.\hat{S}_T.R.\hat{S}_T'.\hat{T}_t^{-}$ with $\hat{T}_t$ & $\hat{S}_T$ as defined respectively in (8.27)-(8.28) and $R = Var\{\underline{\theta}_t/D_{t-1}, V = V_{t-1}\}$.

From the above definition , which coincides with the definition of section 6.3.3 in the case of linear models , we get similar posterior distributions for $\underline{\theta}_t$, $\underline{\mu}_t$ and V as the ones given in 6.3.3 , considering $\hat{F}_t$, $\hat{R}_T$, etc instead of F , R_T , etc . As a consequence , the Updating Equations for the DNLM of this section will be obtained from the one given in section 6.3.4 with some slight adaptations , as shown in the Appendix 8.1 [equations (8.30)–(8.45)] .

8.2.3 - An Application : The Multivariate Seasonal Growth Multiplicative - SGM Model

This model is a natural multivariate extension of the scalar seasonal growth multiplicative model - [Harrison(1965) , Migon(1984)] - and is a special case of the general formulation of section 8.2.1 .

I - Model Definition :

The multivariate seasonal growth multiplicative model for a d-dimensional vector of observations $\underline{y}_t$ has its basic structure described by a linear trend for each individual series and a common multiplicative seasonal component. In full notation , it is defined as follows :

i) the process parameter $\underline{\theta}_t = \begin{pmatrix} \underline{\pi}_t \\ \underline{\rho}_t \end{pmatrix}$ is such that $\underline{\pi}_t = (\underline{\pi}_1, ..., \underline{\pi}_d)_t'$ is a 2dx1 vector accounting for the linear trend components $\underline{\pi}_{.t} = \begin{pmatrix} \mu_{it} \\ \beta_{it} \end{pmatrix}$, i = 1,..,d (level and growth parameters) and $\underline{\rho}_t$ is a 2hx1 vector of seasonal effects (harmonic representation) common for all time series , where h is the number of harmonics considered (or , if the Nyquist frequency is present , this vector is a (2h-1)x1 vector) .

ii) the parameter evolution is expressed by the system matrix

$$\mathbf{G} = diag\,[\,(I_d \otimes G_T)\,,\,(G_{H1}, .., G_{Hh})\,] \tag{8.46}$$

where $G_T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ & $G_{Hj} = \begin{pmatrix} cosjw & sinjw \\ -sinjw & cosjw \end{pmatrix}$ $j = 1, .., h.$ with $w = w_j = \frac{2\pi}{T_j}$ where T_j is the j^{th} seasonal period (For the Nyquist frequency , we have $G_{H,h} = -1$) .

iii) the two basic TS components are expressed by the trend ,

$$\underline{\gamma}_t = \mathbf{F}.\underline{\theta}_t \qquad \text{where} \qquad \mathbf{F} = [\,(I_d \otimes F), \underline{0}\,] \quad \& \quad F = (1,0) \tag{8.47}$$

and the seasonal effects component ,

$$\rho_t = 1 + \mathbf{E}.\underline{\theta}_t \qquad \text{where} \qquad \mathbf{E} = (\underline{0}, 1, 0, .., 1, 0) \tag{8.48}$$

They are linked through the non-linear (multiplicative) relation $g(\underline{\theta}_t) = \rho_t.\underline{\gamma}_t$, which completes the model formulation of equations (8.1)-(8.2).

II - Updating Equations .

As mentioned before, in order to implement the updating equations through the methods of section 8.2.1 , the three quantities given by the equations (8.5)-(8.7) should be evaluated , and this is done considering $\underline{\mu}_t = g(\underline{\theta}_t) = \rho_t.\underline{\gamma}_t$.

Now, since $(\underline{\theta}_t / D_{t-1}, V = I) \simeq N(\underline{a}_t, R_t)$, the joint prior moments for $\underline{\lambda}_t = \begin{pmatrix} \underline{\gamma}_t \\ \rho_t \end{pmatrix}$ will be given by : $(\underline{\lambda}_t / D_{t-1}, V = I) \simeq N(\hat{\underline{\lambda}}_t, \mathbf{R}_t)$, where :

$$\hat{\underline{\lambda}}_t = \begin{pmatrix} \hat{\underline{\gamma}}_t \\ \hat{\rho}_t \end{pmatrix} \qquad \text{and} \qquad \mathbf{R}_t = \begin{pmatrix} V(\underline{\gamma}_t) & CV \\ \cdot & V(\rho_t) \end{pmatrix} \qquad \text{with} \tag{8.49}$$

$$\hat{\underline{\gamma}}_t = \mathbf{F}.\underline{a}_t \qquad \hat{\rho}_t = 1 + \mathbf{E}.\underline{a}_t \qquad V(\underline{\gamma}_t) = \mathbf{F}.R_t.\mathbf{F}_t^T \qquad V(\rho_t) = \mathbf{E}.R_t.\mathbf{E}^T \qquad CV = \mathbf{F}.R_t.\mathbf{E}^T$$

Then , $(\underline{\mu}_t / D_{t-1}, V) \sim (\underline{f}_t, S_c.\Sigma_t.S_c^T)$ where:

$$\text{i)} \quad \underline{f}_t = E[\rho_t.\underline{\gamma}_t / D_{t-1}, .] = \hat{\rho}_t.\hat{\underline{\gamma}}_t + CV \tag{8.50}$$

$$\text{ii)} \quad \Sigma_t = Var[\rho_t.\underline{\gamma}_t / D_{t-1}, V = I] = H_t^T.\mathbf{R}.H_t + \frac{1}{2}.\mathbf{T}_t \tag{8.51}$$

with the following matrix definitions :

$$a) \qquad H_t = \frac{\partial \underline{\mu}_t}{\partial \underline{\lambda}_t}|_{\underline{\lambda}_t} = \begin{pmatrix} Diag(\hat{\rho}_t, .., \hat{\rho}_t) \\ \hat{\gamma}_{1t}, .., \hat{\gamma}_{dt} \end{pmatrix} \tag{8.52}$$

$$b) \qquad \mathbf{T}_t = \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & tr(T_i.\mathbf{R}_t.T_j.\mathbf{R}_t) & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix} \tag{8.53}$$

where $T_{it} = \frac{\partial^2 \mu_{it}}{\partial \underline{\lambda}_t . \partial \underline{\lambda}_t^T} = \begin{pmatrix} \mathbf{O}_{d \times d} & \underline{h}_i^T \\ \underline{h}_i & 0 \end{pmatrix}$ with $\underline{h}_i = (0, ., 1, ., 0) \quad i = 1, ., d$

For a proof of the validity of equations (8.51)-(8.53) see Appendix 8.2 .

Also , the covariance in equation (8.7) can be expressed , as follows :

$$Cov[\underline{\theta}_t, g(\underline{\theta}_t)] = Cov[\rho_t.\underline{\gamma}_t, \underline{\theta}_t] = \hat{\rho}_t.Cov[\underline{\gamma}_t, \underline{\theta}_t] + \hat{\underline{\gamma}}_t.Cov[\rho_t, \underline{\theta}_t]$$

$$\text{Or} , \qquad Cov[\underline{\theta}_t, g(\underline{\theta}_t)] = \hat{\rho}_t.\mathbf{F}.R_t + \hat{\underline{\gamma}}_t.\mathbf{E}.R_t \tag{8.54}$$

$$\text{And also} \qquad A_t^1 = (\hat{\rho}_t.\mathbf{F} + \hat{\underline{\gamma}}_t.\mathbf{E}).R_t.\Sigma_t^{-1} \tag{8.55}$$

8.3 - Modelling Compositional Data .

In principle , the class of models for vector time series analysis and forecasting presented in the last three chapters of this thesis were directed towards the representation of continuous data or , more specifically , the modelling of multivariate normal observations .

In practice however , some real data sets are not formed by continuous observations but discrete ones or representing proportions of a whole as in the case of compositional data for instance .

Such series of proportions or compositions can be obtained (if not directly as the original data) as the result of the division of each component series by the total . The idea behind this initial transformation (as explained by West & Harrison (1989) , chapter 15) is that if general environmental conditions can be assumed to affect each of the series through a common multiplicative factor at each time , then , the convertion to proportions will remove such effects and lead to a simpler analysis .

In this way , in order to enlarge the applicability of our models to such non-normal data , we consider here the use of special transformations in this previously transformed data , in order to have continuous observations again and consequently , be able to use the DLM framework .

Concretly , if $\underline{y}_t$ (t=1,2,..) is a d-dimensional vector time series of positive quantities of similar nature , then , the time series of proportions $\underline{p}_t$ is defined by

$$\underline{p}_t = \left\{ \sum_{j=1}^{d} y_{jt} \right\}^{-1} \cdot \underline{y}_t \tag{8.56}$$

Now , the logistic log-ratio transformation - Aitchison (1986) , can be used to map the vector of proportions $\underline{p}_t$ into a vector of real-valued quantities $\underline{z}_t$, where a particular symmetric version is given by

$$z_{jt} = log \left\{ \frac{p_{jt}}{\hat{p}_t} \right\} = log p_{jt} - log \hat{p}_t \qquad (j = 1, 2, .., d) \tag{8.57}$$

where $\hat{p}_t = \prod_{j=1}^{d} p_{jt}^{\frac{1}{d}}$ is the geometric mean of the elements p_{jt} .

The inverse of this log ratio transformation is the logistic transformation

$$p_{jt} = \frac{exp(z_{jt})}{\sum_{j=1}^{d} exp(z_{jt})} \quad , \quad j = 1, .., d \tag{8.58}$$

Then , if we represent $\underline{z}_t = (z_{1t}, .., z_{dt})'$ as a normal DLM , the observational distribution of the proportions $\underline{p}_t$ will be the multivariate logistic-normal distribution as defined in Aitchison

& Shen(1980) , and a sufficient condition for the vector $\underline{p}_t$ to follow such distribution is that the original series $\underline{y}_t$ is distributed as a multivariate log-normal .

Now , we face another problem : the elements of $\underline{z}_t$ sum to zero by the definition and this implies that V is singular with rank d-1 . This and other singularities , as suggested by West & Harrison(1989) , can be handled as follows . Initially , $\underline{z}_t$ is modelled ignoring the constraint (i.e. , assuming that V is non-singular) and then transforming $\underline{z}_t$ to $L.\underline{z}_t$ where $L = L' = I - d^{-1}.11'$. Then applying the transformation L to the observation equation of the multivariate DLM in $\underline{z}_t$, we get

$$\underline{z}_t^* = F^*.\underline{\theta}_t + \underline{v}_t^* \quad , \quad \underline{v}_t^* \sim N(\underline{0}, V^*) \tag{8.59}$$

$$\underline{\theta}_t = G.\underline{\theta}_{t-1} + \underline{w}_t \quad , \quad \underline{w}_t \sim N(\underline{0}, W) \tag{8.60}$$

where $\underline{z}_t^* = L.\underline{z}_t$, $F^* = L.F$ and $V^* = L.V.L$.

Then , the algorithm of chapter 6 for multivariate DLM analysis is considered for estimating the process parameter and the observational variance V^* . This information is transformed back to obtain V through the relation $V = L^-.V^*.L^-$, where L^- is the MP-generalized inverse of L .

8.4 - Vector Auto-Regressive Dynamic Models

8.4.1 - Introduction :

Since its original introduction by Litterman,R.B.(1980) , the B.V.A.R. (Bayesian Vector Auto-Regressive) method has became popular in the econometric literature as a successful technique for modelling and forecasting multivariate time series . In order to understand its advantages and limitations as well as its relationship with multivariate DLM's , we present here a brief description and analysis of this model .

By Vector Auto-Regression - VAR , we mean a projection (regression) of each element of a vector time-series on its own lags and lags of each of the other elements in the vector . In an econometric jargon , we can say that all variables in the model are endogenous and the whole model structure is in a reduced form . It is supposed that all lag orders (up to a certain order m say) of all variables are present in each equation and the possibility of introducing one deterministic component in each equation (there are as many equations as elements in the vector time series) is also considered .

The VAR model is claimed to be a very general representation with which to approximate the stochastic process generating a multivariate time series , specially if we consider the class of such models with time-varying coefficients . In fact , as shown in the next sub-section , the VAR model (in its dynamic version) can be considered as a particular case of the common components model presented in chapter 5 , depending on the variance specification for its dynamic (random walk type) evolution .

One key feature of such VAR models is the presence of a large number of parameters which if estimated from a not very large data set , will create an overparametrization problem : one finds a very good in-sample fit and a very bad out-of-sample forecast performance .

Classical econometric models (simultaneous equations or structural models) approach this problem relying on econometric theory to suggest which lagged variables should be present or not in each equation . To avoid overfitting , the forecaster is forced to rely heavily on exclusion restrictions , even though that represents an unreasonably strong prior , one that will never be altered by evidence in the data .

An alternative modelling procedure is to consider a large number of lags , say m , for each variable in each equation and impose some sort of parametric constraint , in a deterministic or stochastic way , in order to increase the number of degrees of freedom . A review and discussion of different deterministic and stochastic types of restrictions used in distributed lag models can be found in Young,A.S.(1983) . It is suggested that there are advantages in using the stochastic type of constraints and in a Bayesian framework it is implemented through the use of adequate prior distributions .

Instead of setting lots of coefficients to zero , the BVAR technique specifies through the use of a prior distribution that most coefficients are likely to be close to zero ; the larger the lag of a given variable , the more likely it is that the coefficient is zero . The rationale for this is the fact that more recent values of a variable are more likely to contain useful information about its future movements than older values .

As a simple ilustrative example , consider that the regression coefficients b^i_{jk} (j = 1,..,m) of the j-th lag in each of the d variables (k = 1,..,d) in each of the d equations , with the exception of the first lag , have the following prior distribution ,

$$(b^i_{jk} / D_0) \quad \sim \quad N(0, \alpha.j^{-1}) \tag{8.61}$$

where α is an overall tightness hyper-parameter (in practice its value is set up between 0.1 and

0.3) . The first coefficient b_1 corresponding to the first lag , as suggested by Litterman(1981) , has its initial prior mean set up as 1 in order to represent a random walk type process and an initial prior variance as in (8.61) . A more general set up for the initial priors as the one suggested by the authors is described in Appendix 8.3 . For a more detailed discussion of the BVAR approach , see for instance Litterman(1980,1984) or Doan,Litterman & Sims(1984) .

Now , a brief formal exposition of the BVAR model as well as some analysis and critics is presented in the next sub-section where the relation between such models and the DLM framework is stablished and some natural extensions are proposed .

8.4.2 - BVAR Models : Definition , Critics and Interpretation as a Special DLM .

Definition : An m^{th} order auto-regressive representation for the dx1 vector time series $\underline{y}_t$ is given by

$$\underline{y}_t = \sum_{j=1}^{m} B_j . \underline{y}_{t-j} + \underline{v}_t \quad , \quad \underline{v}_t \sim N(0, V) \tag{8.62}$$

where B_j (j = 1,..,m) are dxd matrices of unknown auto-regressive coefficients , $\underline{y}_t$ is such that any deterministic components (such as constants , trends , seasonals or any exogenous variable) are supposed to be previously eliminated from the data .

The (static) BVAR model is then defined by the observation equation (likelihood) given above plus some initial prior specification for the auto-regressive parameters (prior distribution) like the one given by equation (8.61) for instance . A more general procedure to specify the initial priors is presented in detail in Appendix 8.3 . The prior specification problem in a general multivariate regression context is discussed in Brown,P.J.(1980) and Makelainen,T & Brown,P.J.(1988).

One typical case of a deterministic term that should be previously eliminated from the data is the seasonal component present in many time series . Unfortunately , this important component is supposed to be known or previously estimated by another method , as for instance , using dummy variables as regressors , as suggested by Litterman(1986) .

What is important to notice is the fact that seasonality does not have a stochastic treatment and is not estimated jointly in the model , which is a serious drawback of this BVAR method .

Now , in order to understand better the BVAR model and see more clearly its limitations , we proceed with the following analysis :

From the model definition (static case) , the i^{th} equation (i = 1,..,d) has the following scalar

form :

$$\begin{aligned} y_t^i = & b_{11}^i \cdot y_{1,t-1} + ... + b_{m1}^i \cdot y_{1,t-m} + \\ & b_{12}^i \cdot y_{2,t-1} + ... + b_{m2}^i \cdot y_{2,t-m} + \\ & .....+ \\ & b_{1d}^i \cdot y_{d,t-1} + ... + b_{md}^i \cdot y_{d,t-m} + v_t \end{aligned} \tag{8.63}$$

where b_{jk}^i is the k^{th} element of the i^{th} row of B_j . Or , equivalently

$$y_t^i = (\underline{y}_{t-1}^{'} \quad ,..., \quad \underline{y}_{t-m}^{'}) \cdot \begin{pmatrix} \underline{b}_1^i \\ \vdots \\ \underline{b}_k^i \end{pmatrix} + v_t \tag{8.64}$$

where the $\underline{y}$'s and $\underline{b}^i$'s are respectively the columns of lagged variables and auto-regressive coefficients in (8.62) .

Now , putting together the d equations like (8.63) , we have the following observation equation :

$$\underline{y}_t = (I \otimes F).\underline{\theta}_t + \underline{v}_t \quad , \quad \underline{v}_t \sim N(\underline{0}, V) \tag{8.65}$$

where $F = (\underline{y}_{t-1}^{'} \quad ,.., \quad \underline{y}_{t-k}^{'})$ & $\underline{\theta}_t = vec\left\{ \begin{pmatrix} \underline{b}_1^1 \\ \vdots \\ \underline{b}_m^1 \end{pmatrix} ,.., \begin{pmatrix} \underline{b}_1^n \\ \vdots \\ \underline{b}_m^n \end{pmatrix} \right\}$.

Also , if we consider the possibility of a random walk type variation for $\underline{\theta}_t$, we have the following system equation :

$$\underline{\theta}_t = \underline{\theta}_{t-1} + \underline{w}_t \quad , \quad \underline{w}_t \sim N(\underline{0}, \mathbf{W}_t) \tag{8.66}$$

and the Dynamic BVAR Model is defined by the equations (8.65)-(8.66) together with some prior distribution for $\underline{\theta}_t$ like (8.61) and a prior distribution for V .

It is important to stress the fact that the Dynamic BVAR model is defined independently of the way that $\mathbf{W}_t$ is specified operationaly , which can be done in many different ways . In the computer implementation of the BVAR methodology developed by the authors - the RATS (Regression Analysis of Time Series) package , such operational specification is left to the user . One special way of implementing $\mathbf{W}_t$, which is part of the common component DLM framework of chapter 5 , is to consider

$$\mathbf{W}_t = V \otimes W_t \tag{8.67}$$

where W_t is specified through the use of a vector of discount factors .

As we can see , the equations (8.65)-(8.66) plus the specification in (8.67) define a special or particular case of the common components model presented in chapter 5 , and the so called Dynamic BVAR model , in such circunstances , can be considered as a type of CCM .

As a consequence , this model can carry out not only the advantages of common components such as simplicity and computational efficiency but also its drawbacks (see chapter 5 for a discussion of the limitations of that model) .

In fact , if we redefine the Dynamic BVAR model through the equations (8.65)-(8.67) , and not more through the equations (8.65)-(8.66) as before , we get some important advantages over the traditional formulation . First of all , we can now estimate V jointly with $\underline{\theta}_t$, which did not happen before - see for instance Litterman (1988) . Second and very important , is the fact that now we can introduce other components in the model as for instance a seasonal component with as many harmonics as necessary or exogenous variables as external individual regressors .

This new formulation - the Dynamic BVAR / CCM - has now overcome one of the main drawbacks of the traditional BVAR model : the proper modelling of seasonal data , and also , it has opened the model to the inclusion of exogenous variables when necessary . However , one important question remains : how about the main drawback of the CCM (see chapter 5) , i.e. , the coincidence with individual univariate DLM results .

It may seem a surprise , but it does not happen in the special framework of the dynamic Bayesian vector auto-regressive model and the reason is as follows . Each univariate marginal model - see for instance the equation (8.63) for the i^{th} variable - depends not only on the lagged values of this i^{th} variable , but also on the lagged values of all the others d-1 variables that form the vector time series . In this way , the predictive distribution for the i^{th} variable for instance , depends not only on the past values of this variable but also on the past values of all the other variables , which characterises a general multivariate model for multiple time series .

8.4.3 - Further Extensions to BVAR Models

Now , in order to remove some lasting limitations still present in the Dynamic BVAR / CCM , as for instance , the constraint that the seasonal pathern or exogenous variables (if

present) should be the same (equal number of harmonics or a common regressor) for each individual time series , as well as constraints concerning interventions , we extend the model . For the applications where we need a seasonal component and it can not be represented in a common components framework (or we need different regressors for each marginal equation) , we redefine the Dynamic BVAR model - call that the Extended BVAR model - as a special case of the multivariate DLM model of chapter 6 .

Another case where such general framework (the multivariate DLM) can be used as an extended BVAR model is when we want to remove one or more lagged variables in certain equations of a vector auto-regressive structure . That is , when we do not want or need all lags of all variables present in each equation . This sort of modelling flexibility (not present in the original BVAR model) can be very important in larger systems for instance , where we need to reduce even more the total number of parameters in the model and any useless component should be removed .

Appendix 8.1 - Updating Equations for the DNLM (algorithms I & II)

Using the same notation of chapter 6 , the full Updating Equations for the scaled DNLM will be given by the following steps (algorithm I) :

Notation : $(\underline{\theta}_{t-1} / D_{t-1} , V = I) \sim N(\underline{m}^1_{t-1} , C^1_{t-1})$

step 1 - time updating : $(\underline{\theta}_t / D_{t-1}, V = I) \sim N(\underline{a}_t, R_t)$ where :

$$\underline{a}_t = G.\underline{m}^1_{t-1} \tag{8.8}$$

$$R_t = G.C_{t-1}.G^T + W \tag{8.9}$$

In practice , W is specified through a given vector of discount factors $\underline{b}$. Also , there is no time updating for V , since it is supposed to be constant in time .

step 2 - reparametrization & scaling : $(\underline{\mu}_t / D_{t-1}, V) \sim N(\underline{f}_t , R^*_t)$ where :

$$\underline{f}_t = E\{g(\underline{\theta}_t) / D_{t-1}, V\} \tag{8.10}$$

$$R^*_t = S_c.\Sigma_t.S^T_c \tag{8.11}$$

$$\Sigma_t = Var\{ g(\underline{\theta}_t) / D_{t-1}, V = I\} \tag{8.12}$$

where initially , the scaling matrix S_c is set up as an identity and updated in step 3 .

step 3 - observation updating : $(\underline{\mu}_t / D_t, V) \sim N(\underline{m}_t, C_t)$ where :

$$\underline{m}_t = \underline{f}_t + A_t^* . (\underline{y}_t - \underline{f}_t) \tag{8.13}$$

$$C_t = R_t^* - A_t^* . Q_t^* . A_t^{*'} \tag{8.14}$$

$$A_t^* = R_t^* . Q_t^{*^{-1}} \tag{8.15}$$

$$Q_t^* = R_t^* + V \tag{8.16}$$

Also , $(V / D_t) \sim W^{-1}(d_t ; n_t)$ where :

$$d_t = n_t . V_t = d_{t-1} + h_t . h_t^T \tag{8.17}$$

$$n_t = n_{t-1} + 1 \tag{8.18}$$

$$h_t = S_{t-1} . [(Q_t^*)^{\frac{1}{2}}]^{-1} . \underline{e}_t \tag{8.19}$$

$$S_t = [V_t]^{\frac{1}{2}} \tag{8.20}$$

$$S_c = S_t . S_0 \tag{8.21}$$

where S_0 is a reference matrix set up initialy such that the initial scale factor S_c is the identity matrix .

step 4 - inverse reparametrization and scaling : $(\underline{\theta}_t / D_t , V = I) \sim N(\underline{m}_t^1 , C_t^1)$ where :

$$\underline{m}_t^1 = \underline{a}_t + A_t^1 . (\underline{m}_t - \underline{f}_t) \tag{8.22}$$

$$C_t^1 = R_t + A_t^1 (C_t^* - \Sigma_t) . A_t^{1'} \tag{8.23}$$

$$A_t^1 = Cov\{\underline{\theta}_t , g(\underline{\theta}_t)\} . \Sigma_t^{-1} \tag{8.24}$$

$$C_t^* = S_c^{-1} . C_t . (S_c^{-1})^T \tag{8.25}$$

Updating Equations for the DNLM (algorithm II)

As a consequence of the definitions and explanations of section 8.2.2 , we have the following Updating Equations for the DNLM defined directly in terms of the process parameter $\underline{\theta}_t$:

Notation : $(\underline{\theta}_{t-1} / D_{t-1}, V) \sim N(\underline{m}_{t-1}, C_{t-1})$

i) Time Updating : $(\underline{\theta}_t / D_{t-1}, V) \sim N(\underline{a}_t , R_t)$ where :

$$\underline{a}_t = G . \underline{m}_{t-1} \tag{8.30}$$

$$R_t = G . C_{t-1} . G^T + W_t \tag{8.31}$$

ii) Observation Updating for V : $(V^{-1}/D_t) \sim W(d_t\,;n_t)$ where :

$$d_t = d_{t-1} + h_t.h_t^T \tag{8.32}$$

$$h_t = V_{t-1}^{\frac{1}{2}}.[\hat{Q}_t^{\frac{1}{2}}]^{-1}.\underline{e}_t \tag{8.33}$$

$$\hat{Q}_t = \hat{F}^T.R_t.\hat{F} + V_{t-1} \tag{8.34}$$

$$\underline{e}_t = \underline{y}_t - \hat{F}^T.\underline{a}_t \tag{8.35}$$

$$n_t = n_{t-1} + 1 \tag{8.36}$$

$$V_t = n_t^{-1}.d_t \tag{8.37}$$

$$S_t = V_t^{\frac{1}{2}} \tag{8.38}$$

iii) Scaling and Observational Updating for θ_t : $(\underline{\theta}_t/D_t,V) \sim N(\hat{m}_t\,,\hat{C}_t)$ where :

$$\underline{\hat{m}}_t = \underline{a}_t + \hat{A}_T.\underline{e}_t \tag{8.39}$$

$$\hat{A}_T = \hat{R}_T.\hat{F}.\hat{Q}_T^{-1} \tag{8.40}$$

$$\hat{C}_t = (I - \hat{F}.\hat{A}_T).\hat{R}_T \tag{8.41}$$

$$\hat{R}_T = \hat{T}^-.(\hat{S}_T.R_t.\hat{S}_T).\hat{T}^{-'} \tag{8.42}$$

$$\hat{Q}_T = \hat{F}^T.\hat{R}_T.\hat{F} + V_t \tag{8.43}$$

$$S^* = S_t.S_{t-1} \tag{8.44}$$

$$\hat{S}_T = \begin{pmatrix} S^*.\hat{T}_1 \\ \vdots \\ S^*.\hat{T}_k \end{pmatrix} \tag{8.45}$$

$$\hat{T}_j = \hat{F}'.G^{j-1} \quad , \quad j = 1,.,k \tag{8.45a}$$

Appendix 8.2 - Covariance Between Quadratic Forms & Related Results

i) In order to prove the validity of the equation (8.31) we consider the Taylor Series expansion of the vectorial function $\underline{\mu}_t = \rho_t.\underline{\gamma}_t$ as follows :

$$\underline{\mu}_t = f(\underline{\lambda}_t) = f(\underline{\hat{\lambda}}_t) + H_t^T.(\underline{\lambda}_t - \underline{\hat{\lambda}}_t) + \frac{1}{2}\begin{pmatrix} \vdots \\ (\underline{\lambda}_t - \underline{\hat{\lambda}}_t)^T.T_{it}.(\underline{\lambda}_t - \underline{\hat{\lambda}}_t) \\ \vdots \end{pmatrix}$$

with $H_t = \frac{\partial \underline{\mu}_t}{\partial \underline{\lambda}_t}\Big|_{\underline{\hat{\lambda}}_t} = \begin{pmatrix} \frac{\partial \underline{\mu}_t}{\partial \underline{\gamma}_t} \\ \frac{\partial \underline{\mu}_t}{\partial \rho_t} \end{pmatrix}\Bigg|_{\underline{\hat{\lambda}}_t}$ where:

$$\frac{\partial \underline{\mu}_t}{\partial \underline{\gamma}_t}\Big|_{\underline{\hat{\lambda}}_t} = \frac{\partial \rho_t.(\gamma_{1t},..,\gamma_{dt})}{\partial(\gamma_{1t},..,\gamma_{dt})^T}\Big|_{\underline{\hat{\lambda}}_t} = Diag(\hat{\rho}_t,..,\hat{\rho}_t)$$

$$\frac{\partial \underline{\mu}_t}{\partial \rho_t}\Big|_{\underline{\hat{\lambda}}_t} = \frac{\partial \rho_t.\underline{\gamma}_t}{\partial \rho_t}\Big|_{\underline{\hat{\lambda}}_t} = \underline{\hat{\gamma}}_t^{\,T} = (\hat{\gamma_{1t}},..,\hat{\gamma_{dt}})$$

Also , the 2^{nd} derivatives will form an array where each one of its d faces is given by ,

$$T_{it} = \frac{\partial^2 \mu_{it}}{\partial \underline{\mu}_t \partial \underline{\mu}_t^T}\bigg|_{\underline{\lambda}_t} = \frac{\partial^2 \rho_t \underline{\gamma}_{it}}{\partial(\underline{\gamma}_t^T, \rho_t)\partial(\underline{\gamma}_t^T, \rho_t)^T} = \begin{pmatrix} \mathbf{O}_{dxd} & \underline{h}_i^T \\ \underline{h}_i & 0 \end{pmatrix}$$

with $\underline{h}_i = (0,.,1,.,0)$. Consequently , the variance-covariance matrix of $\underline{\mu}_t$ will be given by,

$$Var(\underline{\mu}_t) = H_t^T.Var(\underline{\lambda}_t).H_t + \frac{1}{4}\begin{pmatrix} . & . & . \\ . & t_{ij} & . \\ . & . & . \end{pmatrix} \qquad \text{where}$$

$$t_{ij} = Cov(\underline{\lambda}_t^{*T}.T_i.\underline{\lambda}_t^*, \underline{\lambda}_t^{*T}.T_j.\underline{\lambda}_t^*) \qquad \text{where} \qquad \underline{\lambda}_t^* = \underline{\lambda}_t - \underline{\hat{\lambda}}_t.$$

which involves covariances between random quadratic forms .

ii) In order to find out an expression for the covariance of two random quadratic forms we proceed as follows . Initially , consider the quadratic forms $\underline{\lambda}^{'}.T_i.\underline{\lambda}$ and $\underline{\lambda}^{'}.T_j.\underline{\lambda}$ where T_i, T_j are symmetric and $\underline{\lambda} \sim (\underline{0}, \mathbf{R})$. As $\mathbf{R}$ is a covariance matrix , it is positive semi-definite , and consequently there exists a non-singular matrix U such that $U.U^{'} = \mathbf{R}$. Then , we can rewrite the quadratic forms as $\underline{\lambda}^{'}.T_i.\underline{\lambda} = X^{'}.A.X$ and $\underline{\lambda}^{'}.T_j.\underline{\lambda} = X^{'}.B.X$ where $X = U^{-1}.\underline{\lambda}$, $A = U^{'}.T_i.U$ and $B = U^{'}.T_j.U$, what gives an uncorrelated structure for X : $X \sim (\underline{0}, I)$. Although no normality assuptions are considered , we suppose independence between the elements X_i of X as well constant 4^{th} order moments ($\alpha_4 = EX_i^4 = \text{const.}$ for all $i = 1,..,d+1$) . Now ,

$$E\{(X^{'}.A.X).(X^{'}.B.X)\} = \sum_{i,j,k,l} a_{ij}.b_{kl}.E\{X_i.X_j.X_k.X_l\}$$

where , using the assumptions of independence and constant 4^{th} order moments , we get

$$E(X_i.X_j.X_k.X_l) = \begin{cases} \alpha_4, & \text{if the indexes are all equal} \\ 1, & \text{if indexes are equal in pairs} \\ 0, & \text{otherwise} \end{cases}$$

Then , if $Q_A = X^{'}.A.X$ and $Q_B = X^{'}.B.X$, we have :

$$\begin{aligned} E\{Q_A.Q_B\} &= \alpha_4.\sum_i a_{ii}.b_{ii} + \sum_{i \neq k} a_{ii}.b_{kk} + 2.\sum_{i \neq j} a_{ij}.bij \\ &= \alpha_4.\underline{a}^{'}.\underline{b} + \sum_{\forall i,k} a_{ii}.b_{kk} - \underline{a}^{'}.\underline{b} + 2.\left\{\sum_{\forall i,j} a_{ij}.b_{ij} - \underline{a}^{'}.\underline{b}\right\} \\ &= (\alpha_4 - 3).\underline{a}^{'}.\underline{b} + (trA).(trB) + 2.tr(A.B) \end{aligned}$$

where $\underline{a}$ & $\underline{b}$ are the column vectors of the diagonal elements of A and B respectivelly . Now, as $E(X^{'}.A.X) = trE(A.XX^{'}) = trA$ and $E(X^{'}.B.X) = trB$, we have :

$$\begin{aligned} Cov\{Q_A, Q_B\} &= E\{Q_A, Q_B\} - E\{Q_A\}.E\{Q_B\} \\ &= \delta + (trA).(trB) + 2.tr(A.B) - (trA).(trB) \\ &= \delta + 2.tr(A.B), \qquad \text{where} \quad \delta = (\alpha_4 - 3).\underline{a}^{'}.\underline{b} \end{aligned}$$

Using the definitions of A and B given before , we have

$$Cov\{\underline{\lambda}^{'}.T_i.\underline{\lambda}, \underline{\lambda}^{'}.T_j.\underline{\lambda}\} = \delta + 2.tr(U^{'}.T_i.UU^{'}.T_j.U)$$

Since $U.U^{'} = \mathbf{R}$, and the trace of a product of matrices is invariant under cyclical permutations , we have finally

$$Cov\{\underline{\lambda}^{'}.T_i.\underline{\lambda}, \underline{\lambda}^{'}.T_j.\underline{\lambda}\} = \delta + 2.tr(\mathbf{R}.T_i.\mathbf{R}.T_j)$$

In particular , if we have normality (or a distribution with the same moments as the normal up to 4^{th} order) then $\delta = 0$, what gives , using our initial notation

$$Cov\{\underline{\lambda}^{*'}.T_i.\underline{\lambda}^{*}, \underline{\lambda}^{*'}.T_j.\underline{\lambda}^{*}\} = 2.tr(T_i.\mathbf{R}.T_j.\mathbf{R})$$

Also , the diagonal elements (variances) are given by

$$Var\{\underline{\lambda}^{*'}.T_i.\underline{\lambda}^{*}\} = 2.tr(T_i.\mathbf{R})^2$$

what is the expression commonly used for such variances in the univariate case .

iii) Finally , we need now to prove that each of the d quadratic forms in the $T_j's$ and the linear form in H present in the Taylor Series expansion are not correlated . This is necessary , in order to garantee that the expression for $V(\underline{\mu}_t)$ does not have terms other than the ones considered . To do this , we proceed as follows . Although we do not assume normality for $Y = \underline{\lambda}^{*}$, we will suppose that Y has a symmetric distribution . Under such conditions , as $EY = E(\underline{\lambda}^{*}) = \underline{0}$, we have

$$Cov\{Y, Y^{'}.T.Y\} = E\{Y.Y^{'}.T.Y\} = \begin{pmatrix} \vdots \\ \sum_{i,j} t_{ij}.E(Y_i.Y_j.Y_k) \\ \vdots \end{pmatrix}$$

and from the symmetry assuptions , as $E\{Y_i.Y_j.Y_k\} = 0$, we have , $Cov\{Y, Y^{'}.T.Y\} = 0$.

Consequently, $Var(\underline{\mu}_t) = H_t^T.\mathbf{R}.H_t + \frac{1}{2}\mathbf{T}_t$ where $\mathbf{T}_t = \begin{pmatrix} . & . & . \\ . & tr(T_i.\mathbf{R}_t.T_j.\mathbf{R}_t) & . \\ . & . & . \end{pmatrix}$ as in (8.13), what complete the proof.

Comments :

i) It is interesting to notice that in the particular case of dimension one (univariate case) , there is only one matrix of second derivatives given by $T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and the matrix H of first derivatives becames a vector $H = \begin{pmatrix} \hat{\rho} \\ \hat{\gamma} \end{pmatrix}$. Also , the variance-covariance matrix of $\underline{\lambda}$ is given by $\mathbf{R} = \begin{pmatrix} V_\gamma & C_{\gamma\rho} \\ . & V_\rho \end{pmatrix}$, and consequently

$$V(\mu_t/D_{t-1}) = (\hat{\rho} \quad \hat{\gamma}).\begin{pmatrix} V_\gamma & C_{\gamma\rho} \\ . & V_\rho \end{pmatrix}.\begin{pmatrix} \hat{\rho} \\ \hat{\gamma} \end{pmatrix} + \frac{1}{2}.tr\left\{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.\begin{pmatrix} V_\gamma & C_{\gamma\rho} \\ . & V_\rho \end{pmatrix}\right\}^2$$

Then , after some easy matrix algebra , we get the following expression for the variance of the product of trend and seasonal components:

$$Var(\rho.\gamma) = \hat{\rho}^2.V_\gamma + \hat{\gamma}^2.V_\rho + 2.\hat{\rho}.\hat{\gamma}.C_{\rho\gamma} + C_{\rho\gamma}^2 + V_\rho.V_\gamma$$

, what coincide with results found in Migon(1984) for the univariate seasonal growth multiplicative model.

ii) Finally, it is important to assess the validity of the assumptions considered when we use results for variances and covariances of quadratic forms . In fact , in the context of section 8.2 , such quadratic forms came from (Taylor Series expansion of) the product of trend and seasonal (random) components . If we know the seasonal effect exactly or very precisely (small variance), the product will behave like a linear function and normality will be a consequence . In practice , we do not know the seasonal effects exactly but its variance is much smaller than the trend variance . Or , in other words, the product of the two components is a non-linear function but it has in practice a close to linear behavior and normality or other assumptions such as the ones we have used will be justified . In fact , in order to obtain expressions for the variance-covariance of $\underline{\mu}_t$ we have considered assumptions weaker than normality , such as constant 4^{th} order moments . But , to implement such expressions in an easier way , we have assumed $\delta = 0$, what means a behavior like the normal up to 4^{th} order moments .

Appendix 8.3 - Initial Priors Seting for BVAR Models

The (static) BVAR model is defined by the likelihood function (observation equation) given by equation (8.62) plus the following prior specification for the auto-regressive parameters :

$$(b^i_{jk} / D_0) \sim N(m(j); V(i,j,k)) \quad i,k = 1,.,d \;\&\; j = 1,.,m. \quad \text{where :}$$

i) b^i_{jk} is the coeffi ient of the j^{th} lag (j = 1,.,m) of the k^{th} variable (k = 1,.,d) in the i^{th} equation (i = 1,.,d) of the linear system defined by (8.62) .

ii) $m(j)$ is the initial prior mean of b^i_{jk} and it is equal zero for all lags j with the exception of the first one where we have m(1) = 1 .

iii) $V(i,j,k)$ is the initial prior variance of b^i_{jk} and is given by the following expression :

$$V(i,j,k) = \alpha.f(i,k).g(j).S_k/S_i \quad \text{where :}$$

α is an overall tightness parameter such that , smaller its value , stronger the constraint is , i.e. , tighter to zero (in practice α is chosen between .1 and .3) .

f(i,k) is an indicator of the influence of the k^{th} variable in the i^{th} equation of (8.62) , where f(i,i) = 1 and $0 < f(i,k) < 1$ for $i \neq k$.

g(j) is an indicator of the relative influence of each lag and is expressed as a decreasing function of the lag order j as for instance , $g(j) = j^{-1}$ (harmonic decreasing) , $g(j) = j^{-2}$ (quadratic decreasing) or $g(j) = c^{j-1}$, $0 \leq c \leq 1$ (geometric decreasing) .

S_k/S_i are coeficients designed to correct possible expected inequalities between observational variances ; for instance , if the k^{th} component of $\underline{y}_t$ has approximately twice the variability of the i^{th} component , then , we should introduce this sort of information through the coeficient $S_k/S_i = 2$.

In fact , in the example given by (8.61) , it was supposed that all variables are equally influential in each equation , that all marginal series have approximately the same variability and that the lag-influence function has an harmonic type of decreasing .

CHAPTER 9

FURTHER MODEL EXTENSIONS AND ILLUSTRATIONS

9.1 - Introduction

Among other issues , this chapter discusses further extensions of the basic methodology for multivariate DLM analysis developed in chapter 6 of this thesis . Such extensions include the introduction of extra flexibility in the model through (random walk type) stochastic changes in the observational variance matrix V , as well as , the introduction of monitoring & intervention facilities to be used whenever judged convenient .

These two special modelling aspects , aimed to improve model flexibility and forecasting performance , are discussed in section 9.2 of this chapter , and further extend the range of applications of the basic model formulation of chapter 6 , initiated in chapter 8 of this thesis .

Finally , section 9.3 of this chapter presents a numerical application concerning a multiplicative seasonal growth model , similar in many aspects to the SGM model of chapter 8 , but considering a linear version of that model (in the log scale) , where the seasonal effect parameters are common to all series . This application , illustrates once more the use of the general techniques proposed in this thesis , mainly in situations where other more restrictive methods for multivariate time series analysis do not succeed .

9.2 - Special Modelling Extensions

9.2.1 - Discounted variance matrix learning

As discussed briefly in chapter 2 of this thesis in the context of univariate DLM's , one simple way of giving extra flexibility to the model is to introduce stochastic changes in the scale parameter V_t , as in Harrison & West(1986) . We borrow such ideas from that context and extend it to apply to the multivariate models of chapter 6 . This is done as follows .

Consider the variance matrix V subject to some random disturbance over the time interval t-1 to t , where such stochastic variation is steady and described by a random walk model for V_t (or some other function of V_t such as V_t^{-1} for instance) . Using the notation of chapter 6 , the precision matrix V_{t-1}^{-1} has a posterior distribution at time t-1 given by $(V_{t-1}^{-1} / D_{t-1}) \sim W(d_{t-1}, n_{t-1})$ where d_{t-1} & n_{t-1} are respectively the shape parameter and the number of degrees of freedom .

The dynamic evolution for the precision matrix V_t^{-1} is then modelled as $V_t^{-1} = V_{t-1}^{-1} + (\delta V^{-1})_t$, where the disturbance term is uncorrelated with (V_{t-1}^{-1}/D_{t-1}) . As a consequence , at time t , we have ,

$$Var\{V_t^{-1} / D_{t-1}\} = Var\{V_{t-1} / D_{t-1}\} + Var\{(\delta V^{-1})_t\}$$

where the first of the two variance components in the right hand side is a function of n_{t-1}^{-1} .

Using the discount concept (see chapter 2) , we may represent the increase in variance given by the last term (disturbance) by a discounted version of the first term , and this is done simply discounting n_{t-1} . The d.f. updating $n_t = n_{t-1} + 1$ of the algorithm of chapter 6 , is substituted by the discounted version $n_t = \beta.n_{t-1} + 1$, where a typical value for the discount factor β is around 0.98 or 0.99 .

The rationale and coherence of such procedures is similar to the corresponding univariate model (more details can be found for instance , in West & Harrison(1989)) .

9.2.2 - Multivariate Monitoring & Intervention

As we have seen in chapter 4 of this thesis , in the context of univariate DLM's , model performance can be assessed sequentially to detect possible structural changes or outliers using Bayes' factors , which constitute the basis for an eventual feed-back intervention in the model . Or , if we have information to anticipate an exceptional occurrence or a future structural change , we can implement feed-forward intervention .

In principle , such ideas can be extended to a multivariate context , since Bayes'factors are defined for multivariate predictive distributions exactly in the same way as before (chapter 4) and with the same properties . However , some important practical questions arise as well as methodological ones . Initially , we have to decide if we assess the model performance of $(\underline{y}_t / D_{t-1})$ jointly or marginaly , and the answer is not trivial .

Concretely , if we have $(\underline{y}_t / D_{t-1}) \sim N(\underline{f}_t , Q_t)$, the straightforward multivariate counterpart of the usual (unidimensional) residual standardisation $y_{it}^* = (y_{it} - f_{it}) / Q_{iit}^{\frac{1}{2}}$ $(i = 1,..,d)$ is given by the square root of the quadratic form $T_t^2 = (\underline{y}_t - \underline{f}_t)'.Q_t^{-1}.(\underline{y}_t - \underline{f}_t)$, where T_t is the so called Hotelling statistic - Hotelling(1947) .

Although used in quality control of industrial processes , a monitoring scheme based on the overall statistic T_t instead of the marginals y_{it}^* would present at least two big inconveniences

for time series applications . First , an assessment measure using Bayes' factors , based on an overall statistic T_t could suggest a good overall model performance concomitantly with one or more predictive marginals having a bad forecasting performance . This unfortunate situation can happen in particular , if the dimension d is large and just one or a few marginals disagree with the current model , since this effect is masked by the overall statistic T_t and a necessary call for intervention in that particular components is not heard .

A second undesirable property of such Bayesian monitoring schemes based on the statistic T_t as a kind of standardised residual is that its operating characteristics , mainly the Expected Run Length - ERL will tend to exibit a slow detection of change because of the masking effect discussed earlier . In fact , the ERL for the detection of level change in normal processes via the marginal and joint approach (in a non-Bayesian quality control context) has been studied by Crosier(1988) , with the results showing the superiority of the marginal approach in relation to the joint one (respectively called multivariate CUSUM and CUSUM of the T statistics by Crosier) . For a brief discussion about the mathematical equivalence between log Bayes factors and the so called CUSUM's , as well as ERL and related topics , see chapter 4 of this thesis .

As a consequence of the discussion and analysis carried above , it is clear that the more appropriate way to extend the Bayesian monitoring scheme of chapter 4 to the multivariate case is to consider a set of d parallel monitoring schemes for each marginal series . Therefore , this is the basis for the extension of automatic feed-back intervention schemes to the case of vector time series . The practical implementation of such schemes , can follow the basic lines described here and in chapter 4 but there is some flexibility to adapt to a particular type of application or objectives and , therefore , this is left to the practitioner .

9.3 - The SGM Model : A Numerical Illustration

9.3.1 - Introduction

As discussed in chapter 5 of this thesis (section 5.3.2) , the modelling of any individual component , as for instance , a common seasonal component , is not compatible with the CCM framework and the need for the general DLM structure is evident . A closely related issue was considered in the previous chapter , concerning the SGM model . In that model , all individual marginal series share the same (multiplicative) seasonal structure , which is what we call a

shared component .

This section considers an alternative version of the SGM model where the observational errors are multiplicative and not additive . In the logarithmic scale , such model may be expressed as a multivariate DLM with trend components for each series but just one seasonal component (with as many harmonics as necessary) common to all marginal series . In principle , both methods are useful to analyse this type of seasonal effect , and one clear advantage of the second one is simplicity . In order to illustrate such modelling aspects , we consider an application with simulated data .

9.3.2 - The Simulated Data : a preliminar view

The data considered here is a three dimensional time series obtained by logarithmic transformation of a vector time series with 48 observations generated according to a SGM model.

The data , shown in fig 9.1 , is typical , for instance , of marketing environments where the sales of one product (corresponding to series 2 in the log scale) falls down concomitantly with the rise in sales of the other two competitors (series 1 and 3) . In such circunstances , where the products are similar and the market is the same , usually , all series are shaped by the same seasonal variations .

From a simple inspection of that picture , the presence of linear growth (trend component) and a 4-period harmonic (seasonal component) is evident . This particular kind of seasonal oscillation is typical of business accounts , since one 'period' is usually 3 months , and 4 periods will correspond to one year .

The presence of a second harmonic in the data , however , is not totaly clear from a simple visual inspection , and the use of a Periodogram is useful to confirm such extra harmonic (this is the Nyquist or 2-period harmonic) .

Once identified the main characteristics in the transformed data that should be considered in the modelling process , we turn to the question of coherence in using transformations of the original data . It should be noticed that the original data (such as sales series , etc) is essentially positive and the logarithmic transformation restores the real domain $\mathbf{R}$ required for analysis purposes (we are using normal DLM's) , and therefore , it is a coherent procedure , as well as , a very usual one .

9.3.3 - The Modelling Structure

The specific DLM $\{F,G,W\}$ to be used in the analysis of the simulated data described in 9.3.2 , is characterised by 3 components (linear trend and two harmonics) with the respective discount factors , as follows :

i) A common real root component with eigenvalue 1 of multiplicity 2 , and a discount factor of 0.95 . (linear trend)

ii) A shared real root component with a single eigenvalue -1 , and a discount factor of 0.975 , related to the series 1,2 & 3 (Nyquist harmonic)

iii) A shared complex root component with eigenvalue of modulus 1 and period 4 , and discount factor of 0.975 , related to the series 1,2 & 3 . (first or 4-period harmonic)

As a consequence , the triple $\{F,G,W\}$ is completely defined (for a review about DLM specification & design , see section 2.4 of chapter 2) . The input of information for the model specification using an interactive APL function is shown after fig 9.1 , as well as the system matrix G . The values chosen for the discount factors represent typical setings and are based on our experience in similar situations (for a discussion about discount factors , see section 2.2 of chapter 2) .

The algorithm for vector DLM analysis (section 6.3 , chapter 6 , using discounted variance learning – section 9.2.1) was applied , considering the model & data set described above , with the following prior set up . As prior mean for the trend (level and growth parameters) and seasonal parameters , was considered respectivelly , 4.5 , 0 , 6 , 0 , 5 , 0 ,1 , 1.2 and 0.5 . The prior observational variance matrix was initialised with 0.8 , 0.8 and 0.4 in the diagonal and zeros elsewhere , with 0.005 for its associated number of d.f. The prior variance matrix was initialised with variances respectively 50 , 10 , 50 , 10 , 50 , 10 , 8 , 8 , 8 , and zeros for the off-diagonal terms . Also , the discount factor for the observational variance learning process was set up as $\beta = 0.99$.

9.3.4 - Estimation and Prediction Results

The one-step-ahead forecasting residuals for the three series are shown in fig 9.2 , at the end of this chapter . Apart from the initial points , the residuals are satisfactory (no large magnitudes or expressive auto-correlations , etc .) , with the most well behaved series (series 3) showing the smallest residuals , as expected .

The estimated correlation structure among series is shown in fig 9.3 . As we can see from

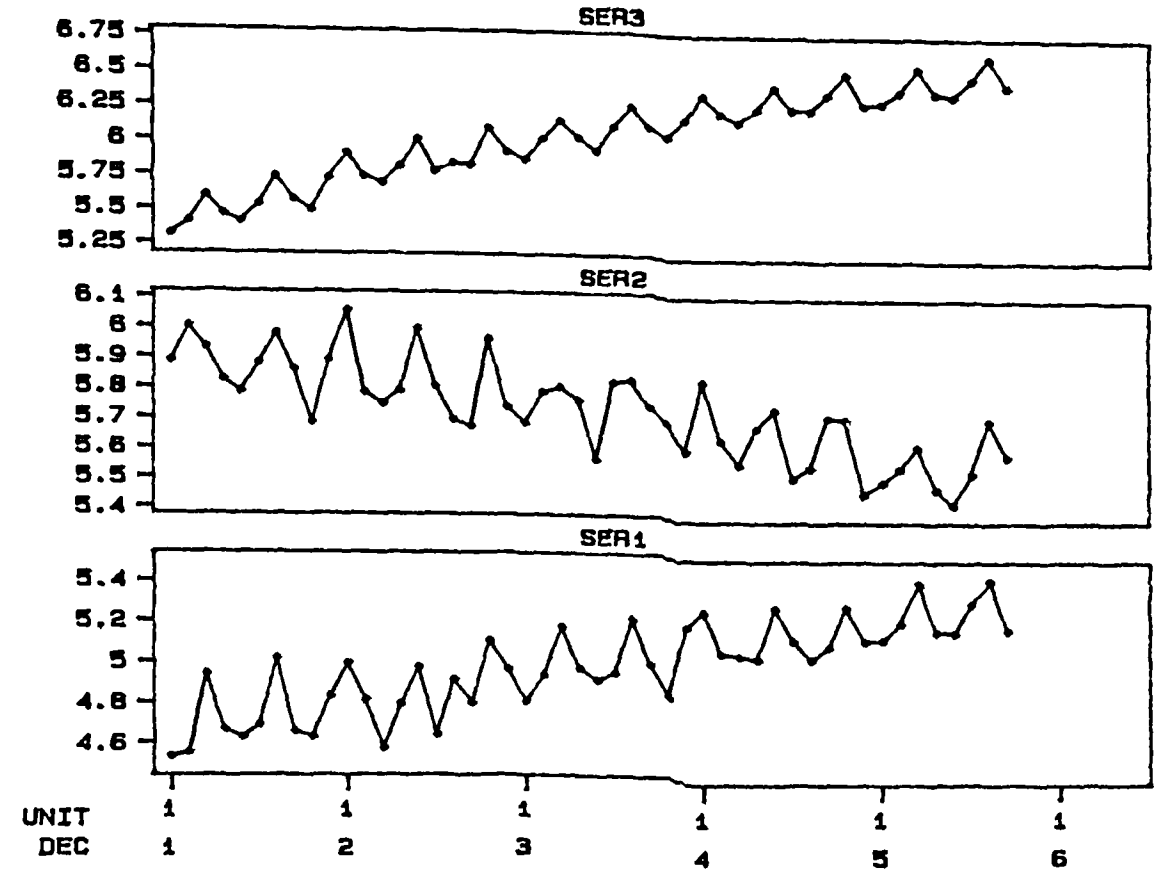
that picture , the learning process for the correlation between the series 2 and 3 , denoted CS[;2;3] , is extremely fast , stabilising around 0.6 with just a few observations . The other two correlations , respectively between series 1 & 3 , and 1 & 2 , have values around 0.3 and -0.3 , with convergence not as quick as in the first picture , but still satisfactory .

The trend component (level parameter) estimates are shown in fig 9.4 , denoted MS[1;] , MS[3;] & MS[5;] , respectively for the series 1 , 2 and 3 . The growth parameter estimates are shown in fig 9.5 , respectively , from the bottom to the top , for series 1 , 2 and 3 . The growth parameter is practicaly constant for the series 1 and 2 , respectively 0.015 and -0.005 , but change (decay) slightly for the series 3 (around 0.025) .

The shared seasonal component estimates are shown in fig 9.6 . The first harmonic , presenting periodicity 4 (the picture on the top) shows a very stable behaviour with time , and a magnitude around 15 per cent . The second harmonic , presenting periodicity 2 (Nyquist frequency) , fluctuates a bit in time , with magnitude around 0.03 , and consequently it is more difficult to detect directly from the data .

Finally , a comment should be made about the estimation of the observational variances , covariances and correlations . The model analysis provides approximate Bayesian estimates of the observational variance-covariance structure , and these estimates are transformed into correlation estimates (intra–series correlations) . The theoretical justification for such procedure is based on the so called plug–in rule (see Quintana,J.M.(1987) , chapter 5) . Very often , estimates of functions of the parameters are of interest , as for instance , the correlation or even the eigen–structure of the matrix V , and such transformations of estimates are themselves Bayesian estimates .

FIG 9.1 SIMULATED VECTOR TIME SERIES WITH COMMON SEASONALITY



Input for Model Specification :

```
DIMENSION of OBSERVATION VECTOR : 3
Do you want any COMMON COMPONENTS : Y
Do you want a common REAL ROOT COMPONENT : Y
MULTIPLICITY : 2
EIGENVALUE : 1
DISCOUNT FACTOR : .95
Do you want a common REAL ROOT COMPONENT : N
do you want a common COMPLEX COMPONENT : N
Do you want any SHARED COMPONENTS : Y
Do you want a REAL ROOT COMPONENT : Y
MULTIPLICITY : 1
EIGENVALUE : -1
Which obs vectors relate : 1  2  3
DISCOUNT FACTOR : .975
Do you want a REAL ROOT COMPONENT : N
do you want a COMPLEX COMPONENT : Y
```

Output : The System Matrix G

1	1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0
0	0	0	1	0	0	0	0	0
0	0	0	0	1	1	0	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	-1	0	0
0	0	0	0	0	0	0	6E-17	1
0	0	0	0	0	0	0	-1	6E-17

FIG 9.2 ONE STEP AHEAD FORECASTING RESIDUALS

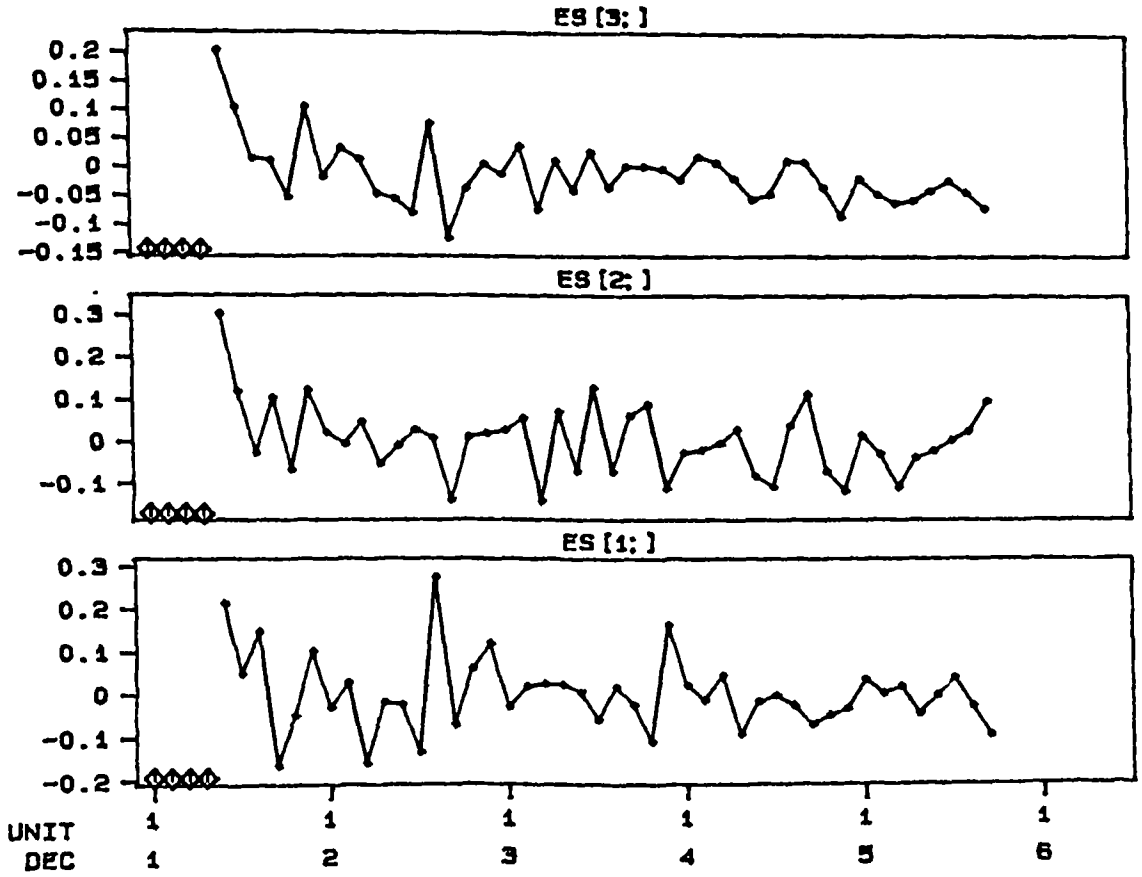


FIG 9.3 CORRELATION BETWEEN SERIES

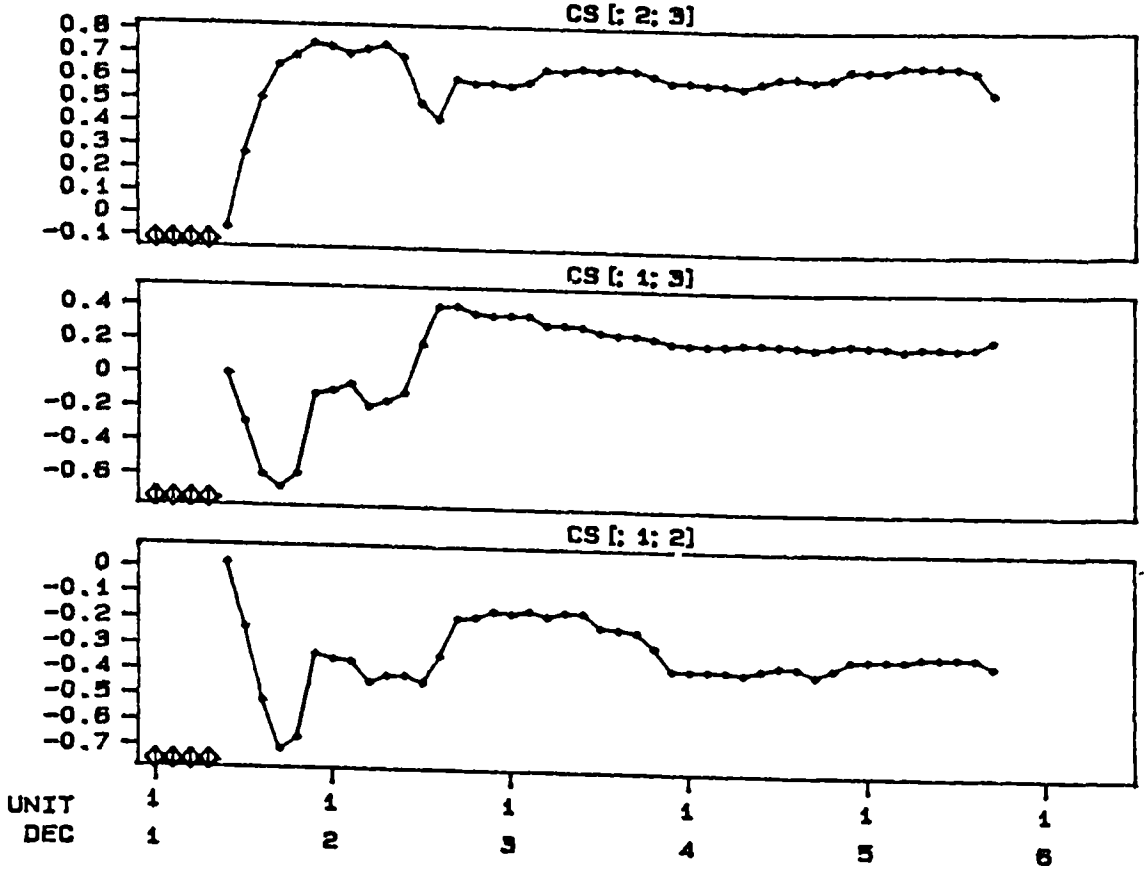


FIG 9.4 TREND COMPONENT ESTIMATES

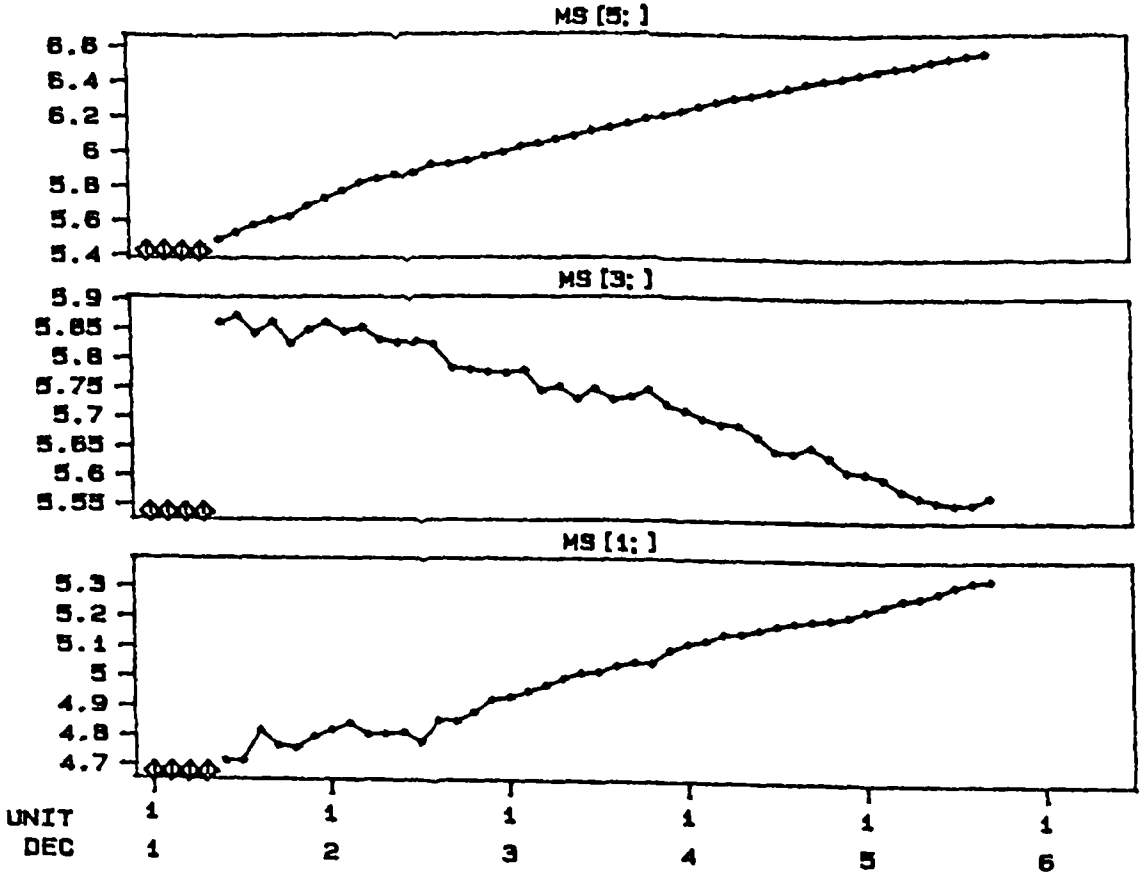


FIG 9.5 GROWTH PARAMETER ESTIMATES

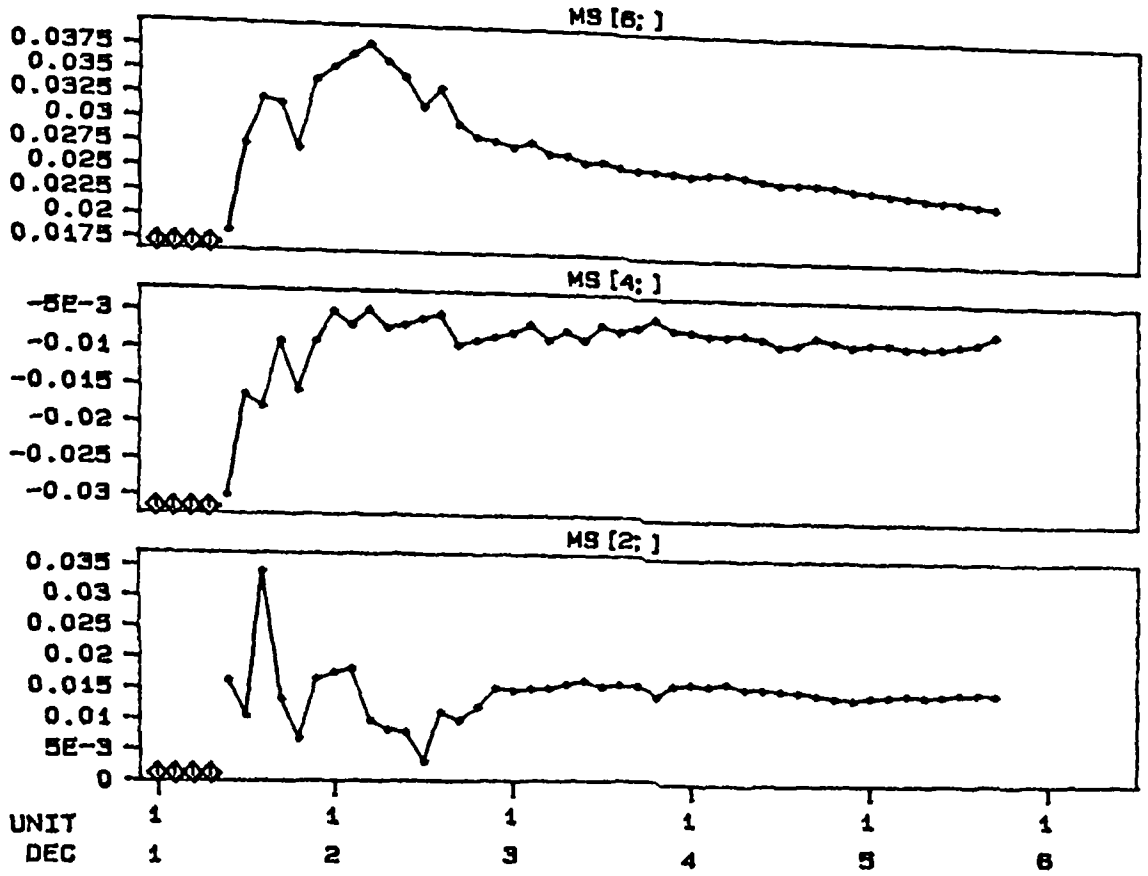
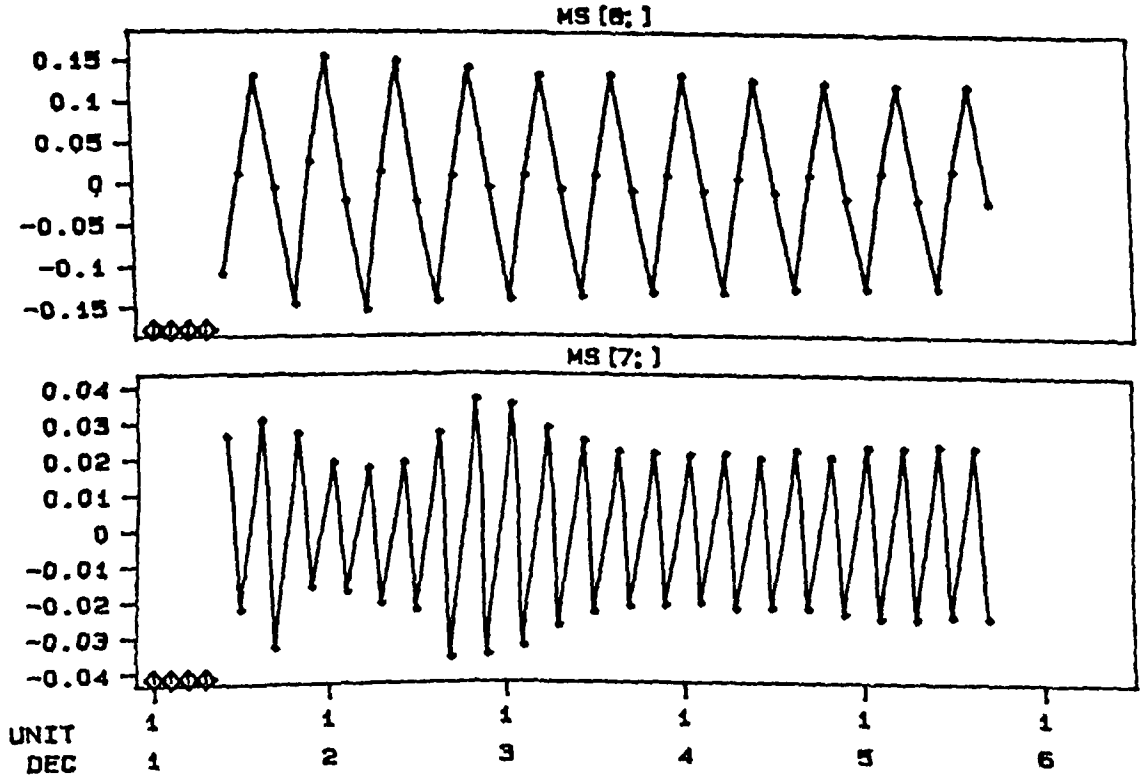


FIG 9.6 COMMON SEASONAL COMPONENT ESTIMATES (2 HARMONICS)



REFERENCES :

Abraham,B. & Ledholter,A.(1983) . *Statistical Methods for Forecasting* . Wiley –Interscience , New York .

Aitchison,J & Dunsmore,I.R.(1980) . *Statistical Prediction Analysis* . Cambridge University Press , Cambridge .

Aitchison,J & Shen,S.M.(1980) . Logistic–normal distributions : some properties and uses . *Biometrika* , **67** , 2 , pp 261–272 .

Aitchison,J.(1986) . *The Statistical Analysis of Compositional Data* . Chapman-Hall , London

Ameen,J.R.M. & Harrison,P.J.(1985) . Normal discount Bayesian models . In *Bayesian Statistics II* , Edited by J.M. Bernardo at al . North– Holland , Amsterdam Univ. Press .

Ameen,J.R.M. & Harrison,P.J.(1985) . Discount Bayesian multi-process modelling with cusums . In *Time Series Analysis : Theory and Practice* **5** , Edited by O.D.Anderson . North-Holland , Amsterdam .

Ameen,J.R.M.(1987) . *Bayesian Sequential Models* . Research Report No. **113** . Department of Statistics , University of Warwick .

Anderson,B.D.O. & Moore,J.B.(1979) . *Optimal Filtering* . Prentice– Hall , New Jersey .

Anderson,T.W.(1984) . *An Introduction to Multivariate Statistical Analysis (second edition)* . New York : Wiley .

Aoki,M.(1967) . *Optimization of Stochastic Systems* . Academic Press , New York .

Aoki,M.(1987) . *State Space Modeling of Time Series* . Springer-Verlag , Berlin .

Barbosa,E.P. & Harrison,P.J.(1989) . Variance Matrix Estimation in Multivariate DLM's . Paper submitted for publication .

Barnard,G.A.(1959) . Control charts and stochastic processes (with discussion) . *J. Roy. Statist. Soc.* **B , 21** , 239–271 .

Bates,D.M. & Watts,D.G.(1980) . Relative Curvature Measures of Nonlinearity (with discussion) . *J. Roy. Statist. Soc.* , Ser **B** , **42**(1) , 1–25 .

Bayes,T.(1763) . An Essay Towards Solving a Problem in the Doctrine of Chances . Reprinted in *Biometrika* , 1958 , **45** , 239-315 .

Berger,J.O.(1985) . *Statistical Decision Theory and Bayesian Analysis* . Cambridge University Press , Cambridge .

Bierman,G.J.(1977) . *Factorization Methods for Discrete Sequential Estimation* . Academic Press , New York .

Bissell,A.F.(1969) . Cusum techniques for quality control . *Appl. Stat.* , **18** .

Box,G.E.P. & Cox,D.R.(1964) . An analysis of transformations (with discussion) . *J. Roy. Statist. Soc.* , Ser **B** , **26** , 241–252 .

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

Box,G.E.P. & Lucas,H.L.(1959) . Design of experiments in non-linear situations . *Biometrika* , **46** , 77–90 .

Box,G.E.P. & Tiao,G.C.(1968) . A Bayesian approach to some outlier problems . *Biometrika* , **55** , 119–129 .

Box,G.E.P. & Tiao,G.C.(1973) . *Bayesian Inference in Statistical Analysis* . Addison-Wesley , Massachusetts .

Broemling,L.D.(1985) . *Bayesian Analysis of Linear Models* . Marcel Dekker , New York .

Brook,D. & Evans,D.A.(1972) . An Approach to the Probability Distribution of CUSUM Run Length . *Biometrika* , **59** , 3 .

Brown,P.J.(1980) . Aspects of Multivariate Regression . *Bayesian Statistics 2* . J.M.Bernardo et al (Eds) . Univ. Press , Valencia , 1980 .

Catlin,D.E.(1989) . *Estimation , Control and the Discrete Kalman Filter* . Springer–Verlag , New York .

Chin,L.(1979) . Advances in adaptive filtering . *Control and Dynamic Systems* C.T. Leondes , Ed. New York : Academic , Vol **15** , pp 278– 356 .

Crosier,R.B.(1988) . Multivariate generalizations of cumulative sum quality control schemes . *Technometrics* , Vol **30** , No. 3 .

Dawid,A.P.(1981) . Some matrix–variate distribution theory : notational considerations and a Bayesian application . *Biometrika* , **68** , 1 , 265–274 .

De Bruyn,C.S.Van Dobben.(1968) .*Cumulative Sum Tests : Theory and Practice* . Griffin and Co. Ltd. , New York .

de Finetti,B.(1974) . *Theory of Probability* . New York : Wiley .

De Groot,M.(1970) . *Optimal Statistical Decisions* . New York : McGraw -Hill .

Doan,T.,Litterman,R. & Sims,C.(1984) . Forecasting and conditional projection using realistic prior distributions . *Econometric Reviews* , **3**(1) , pp. 1–100 .

Downing,D.J. at al.(1978) . Analysis of MUF data using ARIMA models . *J. Inst. Nuclear Mat. Manag.* **7** (4) .

Gamerman,D.(1987) . *Dynamic analysis of survival models and related processes* . Unpublished Ph.D. thesis , Univ. of Warwick .

Gelb,A.(1974) . *Applied Optimal Estimation* . M.I.T. Press , Cambridge .

Goel,A.L. & Wu,S.M.(1971) . Determination of ARL for CUSUM charts . *Technometrics* , Vol **13** , No. 2 .

Goel,A.L. & Wu,S.M.(1973) . Economically optimum design of cusum charts . *Manag. Science* , Vol **19** .

Goldman,A.S. at al.(1982) . Statistical Methods for Nuclear Materials Safeguards : An Overview . *Technometrics* , Vol **24** , No. 4 .

Harrison,P.J. & Davies,O.L.(1964) . The use of CUSUM techniques for the control of routine forecasts . *Oper. Res.* , **12** .

Harrison,P.J.(1965) . Short–term sales forecasting . *Applied Statistics* , **15** , 102–139 .

Harrison,P.J. & Stevens,C.(1971) . A Bayesian approach to short-term forecasting . *Oper. Res. Quart.* , **22** , 341-362 .

Harrison,P.J. & Stevens,C.(1976) . Bayesian forecasting (with discussion) . *J. Roy. Statist. Soc.* , **B** , **38** , 205-247.

Harrison,P.J.(1985) . Convergence for Dynamic Linear Models . Research Report No. 67 , Department of Statistics , University of Warwick .

Harrison,P.J. & West,M.(1987) . Practical Bayesian forecasting . *The Statistician* , **36** , 115-125.

Hartigan,J.A.(1969) . Linear Bayes methods . *J. Roy. Statist. Soc.* , **B** , **31** , 446-454.

Harvey,A.C.(1981) . *Time Series Models* . Philip Allan , Oxford .

Harvey,A.C.(1986) . Analysis and generalization of a multivariate exponential smoothing model . *Management Science* , **32** , 374-380.

Ho,Y.C. & Lee,R.C.K.(1964) . A Bayesian approach to problems in stochastic estimation and control , *I.E.E.E. Trans. Automat. Control* , **AC** , **95** , 333-339 .

Hotelling,H.(1947) . Multivariate Quality Control , *Techniques of Statistical Analysis* , Eisenhart , Hastay and Wallis Editors . McGraw–Hill , New York .

Jeffreys,H.(1961) . *Theory of Probability* (third edition) . Oxford : Claredon Press .

Kalman,R.E.(1960) . A new approach to linear filtering and prediction problems . *J. of Basic Engineering* , **82** , 35–45 .

Kemp,K.W.(1958) . Formulae for calculating the operating characteristics of some sequential tests . *J. Roy. Statist. Soc* , **B** .

Kemp,K.W.(1971) . Formal expressions which can be applied to cusum charts . *J. Roy. Statist. Soc.* , **B** , **33** , No. 3 .

Laplace,P.S.(1774) . Memoir on the Probability of the causes of events . Reprinted in translation in *Statistical Science* , 1986 , Vol 1 , No. 3 , 359-378 .

Lindley,D.V.(1965) . *Introduction to Probability and Statistics* . Cambridge : Cambridge University Press .

Litterman,R.B.(1980) . A Bayesian procedure for forecasting with vector auto–regressions . *Working Paper* . M.I.T. Available from the Research Department , Federal Reserve Bank of Minneapolis .

Litterman,R.B.(1986) . Forecasting with Bayesian Vector Autoregressions : Five Years of Experience .*J. of Busines & Econ. Statist.* , Vol **4** , No. 1 .

Magnus,J.R. & Neudecker,D.(1988) . *Matrix Differential Calculus with Applications in Statistics and Econometrics* . New York : Wiley .

Makelainen,T & Brown,P.J.(1988) . Coherent Priors for Ordered Regressions . *Bayesian Statistics 3* , pp. 667–696 . J.M.Bernardo et al (Eds) . Oxford Univ. Press , 1988 .

Makov,U.E.(1988) . On Stochastic Approximation and Bayes Linear Estimators . *Bayesian Statistics 3* , pp. 697–699 . J.M.Bernardo et al (Eds) . Oxford Univ. Press , 1988 .

Masreliez,C.J.(1975) . Approximate non-Gaussian filtering . I.E.E.E. Trans. Automat. Control , **20** , 107-110.

Maybeck,P.S.(1982) . *Stochastic Models , Estimation and Control* , (Vols 1,2,3) . Academic Press , New York .

Mehra,R.K.(1972) . Approaches to Adaptive Filtering . *I.E.E.E. Trans. Automat. Control* , Vol **AC – 17** , pp 693–698 .

Migon,H.S.(1984) . *An approach to non-linear Bayesian forecasting problems with applications* . Unpublished Ph.D. thesis , Univ. of Warwick .

O'Hagan,A.(1988) . *Probability : Methods and measurement* . Chapman and Hall , London .

Page,E.S.(1954) . Continuous Inspections Schemes . *Biometrika* , **41** .

Plackett,R.L.(1950) . Some theorems in least squares , *Biometrika* **37** , 149-157 .

Pole,A. & West,M.(1987) . Reference analysis of the DLM . *Research Report* , **119** , Dep. of Statistics , Univ. of Warwick .

Pole,A.& West,M(1988) . Efficient numerical integration in dynamic models . *Research Report* **136** , Dep. of Statistics , Univ. of Warwick.

Press,S.J.(1982) . *Applied Multivariate Analysis : Using Bayesian and Frequentist Methods of Inference* . Krieger , Florida .

Press,S.J.(1989) . *Bayesian Statistics* . New York : Wiley .

Press,W.H. at al.(1986) . *Numerical Recipes* . Cambridge Univ. Press , Cambridge .

Quintana,J.M.(1985) . A dynamic linear matrix-variate regression model . *Research Report* **83** , Dep. of Statistics , Univ. of Warwick .

Quintana,J.M.(1987) . Multivariate Bayesian forecasting models . *Unpublished Ph.D. thesis* , University of Warwick .

Quintana,J.M. & West,M.(1987) . Multivariate time series analysis . *The Statistician* , **36** , 275–281 .

Ramsay,J.O. & Novick,M.R.(1980) . PLU Robust Bayesian Decision Theory : Point Estimation . *J. Amer. Statist. Assoc.* , Vol **75** , No. 372 .

Rao,C.R.(1973) . *Linear Statistical Inference and Its Applications* (second edition) . New York : Wiley .

Savage,L.J.(1962) . *The Foundations of Statistics* (second edition) . New York : Wiley .

Seber,G.A.(1977) . *Linear Regression Analysis* . New York : Wiley .

Shaw,J.E.H.(1987) . Aspects of numerical integration and summarisation . In *Bayesian Statistics III* , Edited by J.M.Bernardo at al . Oxford Univ. Press .

Smith,A F.M. at all (1985) . The implementation of the Bayesian paradigm . *Commun. Statist. : Theor. Meth.* , **14** , 1079-1102.

Smith,A.F.M. & Pettit,L.I.(1985) . Outliers and influential observations in linear models . In *Bayesian Statistics II* , Edited by J.M. Bernardo at al. North–Holand , Amsterdam Univ. Press .

Smith,J.Q.(1987) . *Decision Analysis : A Bayesian Approach* . Chapman and Hall , London .

Sorenson,H.W. & Alspach,D.L.(1971) . Recursive Bayesian estimation using Gaussian sums . *Automatica* , **7** , 465–479 .

Speed,T. & Chulpin,D.(1986) . The role of Statistics in N.M.A. : Issues and Problems . *J. Roy. Styatist. Soc.* , Ser **A** , Vol **149** , No. 4 .

Tiao,G.C. & Tsay,R.S.(1989) . Model Specification in Multivariate Time Series . *J. Roy. Statist. Soc.* , Ser. **B** , **51** , No. 2 .

Wald,A.(1947) . *Sequential Analysis* . New York : Wiley .

West,M.(1981) . Robust sequential approximate Bayesian estimation . *J. Roy. Statist. Soc. - B* , **43** , 157-166.

West,M.(1982) . Aspects of recursive Bayesian estimation . *Unpublished Ph.D. thesis* , University of Nottingham .

West,M.(1986) . Bayesian model monitoring . *J.Roy.Statist. Soc.* **B** , **48** .

West,M & Harrison,P.J.(1986) . Monitoring and adaptation in Bayesian forecasting models . *J. Amer. Statist. Assoc.*

West,M & Harrison,P.J.(1989) . *Bayesian Forecasting and Dynamic Models* . . New York : Springer-Verlag .

Young,A.S.(1983) . A Comparative Analysis of Prior Families for Distributed Lags . *Empirical Economics* , Vol **8** , pp 215–227 .

Zellner,A.(1971) . *An Introduction to Bayesian Inference in Econometrics* . . New York : Wiley .

Zellner,A.(1978) . Jeffreys–Bayes Posterior Odds Ratio and the Akaike Information Criteria for Discriminating between Models . *Economic Letters* , **1** , pp. 337–342 .