

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

A Thesis Submitted for the Degree of PhD at the University of Warwick

<http://go.warwick.ac.uk/wrap/34805>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

MULTIVARIATE/BAYESIAN FORECASTING MODELS

JOSÉ MARIO QUINTANA

Thesis submitted for the degree of Doctor of Philosophy

**DEPARTMENT OF STATISTICS
UNIVERSITY OF WARWICK
MARCH 1987**

CONTENTS

Summary vii

CHAPTER 1. INTRODUCTION

1.1 Multivariate Forecasting Models	1
1.2 The Bayesian Approach	2
1.3 Tractability	3
1.4 Outline of the Thesis	4
1.5 Notation	6

CHAPTER 2. THE DYNAMIC LINEAR MODEL

2.1 Model Formulation	7
2.2 Update Computations	8
2.3 Standard Models	9
2.4 Multivariate Models	11
2.5 Specification of the Variances	11

Appendix A2.1 Polynomial Transition Matrix 14

CHAPTER 3. DYNAMIC LINEAR MATRIX-VARIATE REGRESSION

3.1 Model Formulation	15
3.2 Update Computations	18
3.3 The DLMR as a DLM	20
3.4 Multi-process Models	23
3.5 Reparameterization	27
3.6 System Decoupling	28
3.7 Dynamic Weighted Multivariate Regression	29

Appendix A3.1 Kronecker Product and vec Operator 32

Appendix A3.2 Basic Matrix-variate Distribution Theory 35

CHAPTER 4. IMPLEMENTATION ASPECTS

4.1 Implementation via State-space Filters	43
4.2 Implementation via the Sweep Operator	46

Appendix A4.1 The Sweep Operator 50

CHAPTER 5. PLUG-IN ESTIMATION, INFORMATION AND DYNAMIC LINEAR MODELS

5.1 Dynamic Linear Models	56
5.2 Example: Energy Consumption by Primary Fuel Inputs	62
Appendix A5.1 Plug-in Estimation and Information	74
Appendix A5.2 Entropy and Information of Useful Matrix-variate Distributions	79

CHAPTER 6. DYNAMIC RECURSIVE MODEL AND DYNAMIC SCALE VARIANCE

6.1 Dynamic Recursive Model	81
6.2 Example: Hierarchical Missing Observations	85
6.3 Dynamic Scale Variance	88
6.4 Example: Exchange Rate Dynamics	91
Appendix A6.1 Simulation	104
Appendix A6.2 Distribution of Swept Matrices	106

CHAPTER 7. MODELLING ASPECTS

7.1 Vague Priors	108
7.2 Transformations	108
7.3 Special Models	110

CHAPTER 8. DISCUSSION AND FURTHER RESEARCH

References	117
-----------------------------	------------

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Mike West. His excellent guidance, encouragement and friendship made these years of research a very pleasant experience. Also, I am grateful to my fellow students and the members of the staff. In particular, Jo Crosse made my relationship with *AMS-TEX* a lot easier, discussions and lessons with Jeff Harrison were illuminating, Tony O'Hagan kindly offered me a copy of his useful personal notes on matrix-variate distributions and pointed out many mistakes in an earlier version. Special thanks to Federico O'Reilly; he not only encouraged me to take these steps, but to a great extent made it possible. My wife Carol in a recent conversation with her mother said "... and soon we will submit the thesis." If anything, she was being modest.

I am indebted to CONACYT, México for financial support. Material from Crown-copyright records made available through the Central Statistical Office and the SRCC Data Archive has been used by permission of the Controller of H.M. Stationery Office.

To my wife Carol, and our children Quetzal, Chel and ...

The astrologer and sorcerer-fortuneteller brought out the Book of the Horoscope, together with the calendar.

Fray Diego Durán - The Ancient Calendar

SUMMARY

This thesis concerns theoretical and practical Bayesian modelling of multivariate time series. Our main goal is to introduce useful, flexible and tractable multivariate forecasting models and provide the necessary theory for their practical implementation.

After a brief review of the dynamic linear model we formulate a new matrix-variate generalization in which a significant part of the variance-covariance structure is unknown. And a new general algorithm, based on the sweep operator is provided for its recursive implementation. This enables important advances to be made in long-standing problems related with the specification of the variances. We address the problem of plug-in estimation and apply our results in the context of dynamic linear models. We extend our matrix-variate model by considering the unknown part of the variance-covariance structure to be dynamic. Furthermore, we formulate the dynamic recursive model which is a general counterpart of fully recursive econometric models. The latter part of the dissertation is devoted to modelling aspects. The usefulness of the methods proposed is illustrated with several examples involving real and simulated data.

CHAPTER 1

INTRODUCTION

The use of conditional probability as the basis for statistical analysis can be traced back to the work of Bayes (1763) in the eighteenth century. At the beginning of the following century Legendre and Gauss published the method of linear least-squares that they developed working independently of each other (Plackett, 1972). In this century, after a dark period, there has been a revival of Bayesian ideas lead by de Finetti and others (Houle, 1983). Meanwhile, Plackett (1950) obtained a recursive solution for linear least-squares, and Kalman and others (circa 1960) using state-space formulations designed optimal recursive filters for the estimation of multivariate stochastic dynamic linear systems (Gelb, 1974). It soon became apparent that the Bayesian approach provided a neat theoretical framework for the recursive estimation and control of stochastic systems (Ho and Lee, 1964; Aoki, 1967). The merits of this kind of approach for time series and forecasting were evident with the introduction of the Dynamic Linear Model of Harrison and Stevens (1976). This reformulation of the state-space models furnished a time invariant interpretation of the system parameters, model building from simple components, multi-process models, intervention analysis, etc.; a milestone that has stimulated a line of research leading to a methodology known as Bayesian forecasting. This is the base upon which we build our multivariate Bayesian forecasting models.

1.1 MULTIVARIATE FORECASTING MODELS.

We are concerned with Bayesian multivariate forecasting models. First, we wish to stress what we mean by a model. It is often assumed that there is a true underlying model, a set of fundamental "laws", which generates the observations in the real world and our task is to discover it. However, it has been argued that no model used in practice is perfect (Maybeck, 1979, 1982a). Furthermore, the uniqueness and even the existence of a true model will always be open to question (Dickey, 1976; Dawid, 1986). Therefore, instead of searching for a utopia, we adopt a pragmatic approach in line with Harrison and Stevens (1976): a model represents the way in which the observer looks at the observations and their context. In addition, the models entertained are probabilistic, dynamic and Bayesian. The observer does not have a procedure for predicting with perfect accuracy the following observations, and the structure of the model itself is changing in an uncertain way as time passes. Thus, uncertainty is incorporated in the observer's model, the observer is always open-minded and changes his/her mind by means of Bayes' theorem as new information becomes available.

Multivariate forecasting models deal simultaneously with several univariate time series denoted by y_{it} ($i = 1, \dots, q$, $t = 1, 2, \dots$), where t is the time index of equally spaced observations, and q is the number of components observed at any given time. The main goal is to produce joint predictive distributions of sets $y_{t+s} = (y_1, \dots, y_q)_{t+s}$ for $s \geq 0$ using the information available at time $t - 1$.

The essential reason for considering joint multivariate time series as opposed to several marginal time series is that, apart from a trivial case, the marginal predictive distributions do not provide enough information in order to produce a proper joint predictive distribution. Furthermore, the contemporaneous joint variation of the observations is originally uncertain and a main goal of the time series analysis is to learn about it. In so doing, the solution of practical decision problems, which depends on the joint variation, is possible. Moreover, the joint predictive distribution provides a means of updating the predictive distribution of a subset of observations when another subset of contemporaneous observations is given, or more generally, when information regarding another subset of contemporaneous observations is given. Thus, the forecasting performance for univariate time series may be improved by looking at them together with several related time series.

1.2 THE BAYESIAN APPROACH.

The models considered in this dissertation are Bayesian. Although the use of the Bayesian paradigm in statistics is still somewhat controversial, our approach is justified if only for its simplicity. Instead of working with an endless list of ad hoc non-Bayesian procedures, in Bayesian inference we only need to make use of formal golden rules of probability for learning, prediction, etc.

The analysis of Bayesian dynamic models is in principle essentially as neat as the usual static Bayesian analysis. Let y be a set (scalar, vector or matrix) of observations and θ be a set (scalar, vector or matrix) of parameters. The standard static model is defined by the observational density conditional on the parameters (likelihood) $p(y|\theta)$ and the prior density $p(\theta)$, then the predictive density $p(y)$ and the posterior density $p(\theta|y)$ are obtained by means of the conglomerative property and Bayes' theorem,

$$p(y) = \int_{\theta} p(y|\theta) p(\theta) d\theta \quad \text{and} \quad p(\theta|y) = \frac{p(y|\theta) p(\theta)}{p(y)}. \quad (1.1)$$

A lucid and brief exposition of the basic principles can be found in Zellner (1971, Chapter 2).

In a dynamic model the observational density conditional on the set of parameters, also called the system state, is $p(y_t|\theta_t)$, where the state θ_t is changing as time passes. The evolutional density which describes a Markovian transition from the state θ_{t-1} to θ_t is $p(\theta_t|\theta_{t-1})$ and the prior information about θ_{t-1} is represented by $p(\theta_{t-1})$. It is also implicitly assumed that these densities are conditional on the history of the series up to time $t - 1$; see Section 1.5.

It is evident that a dynamic model can accommodate easily a static model. Conversely, a dynamic model can be thought of as static at time t by taking $y = y_t, \theta = (\theta_t, \theta_{t-1}), p(y|\theta) = p(y_t|\theta_t), p(\theta) = p(\theta_t|\theta_{t-1}) p(\theta_{t-1})$. The cycle can be repeated as often as necessary since the next prior for θ_t , given the past and present information, is given by $p(\theta_t|y_t) = \int_{\theta_{t-1}} p(\theta_t, \theta_{t-1}|y) d\theta_{t-1}$. In summary, the dynamic model can be seen as static by considering jointly the parameters θ_t and θ_{t-1} . Their joint prior information is expressed in terms of the evolutional information and the prior information about θ_{t-1} . The set of parameters θ_{t-1} do not appear in the observational density for y_t , but this is irrelevant to the analysis. This representation is employed in Chapter 3 in order to formulate a general multivariate dynamic linear model.

A major criticism of the Bayesian viewpoint has been its lack of objectivity (misleadingly associated with the prior information). Nevertheless, there is no point in arguing about whether or not the non-Bayesian approaches are objective, since the meaning of objectivity itself is, in our opinion, very subjective! Instead of wasting more space discussing the Bayesian - non-Bayesian controversy, we refer the reader to the extensive literature on this recurrent topic: Berger (1985), Barnett (1982), de Finetti (1974, 1975), Lindley (1971), Dempster (1969), etc.

1.3 TRACTABILITY.

It is easy to see that, in the seemingly static representation, the key calculations needed for a dynamic model are simply:

$$p(\theta_t) = \int_{\theta_{t-1}} p(\theta_t | \theta_{t-1}) p(\theta_{t-1}) d\theta_{t-1}, \quad (1.2a)$$

$$p(y_t) = \int_{\theta_t} p(y_t | \theta_t) p(\theta_t) d\theta_t, \quad \text{and} \quad (1.2b)$$

$$p(\theta_t | y_t) = \frac{p(y_t | \theta_t) p(\theta_t)}{p(y_t)}. \quad (1.2c)$$

Equation (1.2a) gives the density of θ at time t in terms of the evolutional density and the previous density of θ at time $t-1$ and represents our knowledge of θ at time t before observing y_t . The one-step ahead predictive density is provided by (1.2b). Finally, (1.2c) closes the cycle since it is the formula for updating the density of θ_t after observing y_t . Long-term forecasting is achieved by applying (1.2a) and (1.2b) repeatedly, but, of course, skipping the updating formula (1.2c).

Successful implementation of equations (1.2) depends on the difficulties encountered in solving the integrals appearing in (1.2a) and (1.2b). Perhaps a not so obvious and even more critical factor, as pointed out by Aoki (1967, Appendix 4), is the existence of a tractable sufficient statistic of fixed dimension for θ_t . If such a tractable statistic exists then the complexity of the density of θ_t does not grow as time passes, and so is updated easily. A statistic is simply a function of the observations and a sufficient statistic for a set of parameters, by definition, summarizes all the information that the observations provide about the parameters. In other words, the distribution of the parameters conditioning on the observations is the same as the distribution obtained by conditioning only on the sufficient statistic. By a tractable statistic we mean a function of the observations which is easily computated. A sufficient statistic of fixed dimension always can be constructed, in principle, with the aid of a bijective function from the real line to the plane, e.g. Simmons (1963, p. 37-38). However, it does not have any practical use since its implementation requires a computer with infinite word-length and infinite speed!

The discussion above suggests that the difficulties in the implementation of (1.2) can be minimized when the marginals of the joint density $p(\theta_{t-1}, \theta_t)$ belong to a family with hyperparameters of fixed dimensions. In addition, marginal densities must be closed under sampling (self-reproducing) relative to the likelihood. Finally, the posterior hyperparameters must be tractable functions of the observations

and the prior hyperparameters. Thus, it is not surprising that a good deal of attention has been paid to dynamic normal models (with known variance-covariance structure), e.g. Aoki (1967), Harrison and Stevens (1971, 1976), Maybeck (1979, 1982a, 1982b), etc.

The major aim of this thesis is to extend as much as possible the dynamic normal models to the multivariate case in which a significant part of the variance-covariance structure is unknown, but to retain their tractability so that they can be implemented efficiently in a typical personal computer.

1.4 OUTLINE OF THE THESIS.

The Dynamic Linear Model (DLM) is reviewed in Chapter 2. This includes model formulation; updating formulas for the posterior hyperparameters in terms of the prior hyperparameter and the observations; the use of the superposition principle for building models with polynomial, harmonic and/or damped trends; multivariate DLM models; and the problems of specifying observational and evolutional variances. For polynomial trends we use a setting in terms of powers rather than the conventional one in terms of standarized factorial polynomials. The verification of this alternative polynomial trend setting is provided in Appendix A1.1.

Chapter 3 is the theoretical core of the thesis. An extension of the DLM, the Dynamic Linear Matrix-variate Regression (DLMR) model is developed. In this new model an important part of the variance-covariance structure, the scale variance matrix, is assumed unknown with an inverted-Wishart distribution. In so doing, we make significant advances in long-standing problems concerning the specification of the variance-covariance structure in DLM's. As a row vector the DLMR is a Dynamic Weighted Multivariate Regression (DWMR) and as a column vector it is a multivariate DLM. In Section 3.1 we formulate the DLMR model. Its update computations are derived in Section 3.2. These include a recurrence for updating the inverted-Wishart hyperparameters associated with the scale variance matrix. In Section 3.3 we reformulate the DLMR in a DLM form. Using this representation we show how modelling with the DLMR is very similar to modelling with the DLM. In addition, by employing this form we interpret the scale variance matrix as a (Kronecker) scale factor of the observational and evolutional variances. In Section 3.4 we generalize the DLM Multi-process models (Harrison and Stevens, 1971, 1976) in order to cope with the DLMR. In particular, we obtain new collapsing formulas for multi-process models Class II. In Section 3.5 we discuss equivalent reparameterizations of DLMR's. Decoupling a DLMR into several independent DLMR's is the subject of Section 3.6. A testing procedure based on the Jeffreys' technique is provided. We close the chapter with a discussion of the DWMR model. This useful model contains as a special, static case the usual Weighted Multivariate Regression (WMR) model. Hence, it offers a general and more realistic alternative to the WMR for modelling multivariate time series. This situation is the multivariate analogy to that of the univariate DLM and the multiple linear regression. For the non-weighted DWMR the scale variance matrix can be regarded as the observational variance. Therefore, we have, in this case, an effective on-line observational variance learning procedure. Two appendices are provided. The relevant results about the Kronecker

product and vec operator are given in Appendix A3.1. The basic Matrix-variate Distribution theory is developed in Appendix A3.2. Special attention has been paid to including both singular and non-singular distributions.

The implementation of the DLMR updating recurrences is the theme of Chapter 4. Time spent on this topic is worthwhile since the DLMR contains several dynamic and static models as special cases. In Section 4.1 we generalize the Kalman, Joseph, Square-root and Inverse Covariance state-space filters (Maybeck, 1979). In particular, we obtain new recurrences for the hyperparameters associated with the scale variance matrix. A filter based on the Efroymson (1960) sweep operator is introduced in Section 4.2. The usual assumption about non-singularity is dropped. Hence, this new filter provides the most general implementation possible and yet it is surprisingly easy to put into practice. The essential sweep operator theory is given in Appendix A4.1. In particular, a key separation principle is identified.

Prediction via plug-in estimation in the context of dynamic linear models is discussed in Chapter 5. The loss function associated with the plug-in estimators (PIE's) is the Kullback and Liebler (1951) directed divergence between the actual (unknown) likelihood and the (plug-in) estimated likelihood. In Section 5.1 we derive the PIE's for the DLMR. We evaluate the cost of using an estimated observation distribution instead of the predictive distribution for forecasting purposes. The PIE's have an appealing property: they are invariant under one-to-one parametric transformations. In Section 5.2 the use of PIE's as point estimators is explored and illustrated with an example involving energy consumption data. In Appendix A5.1 we undertake the problem of plug-in estimation from a decision theoretic viewpoint. The general results obtained constitute the basis for the material presented in Section 5.1. In Appendix A5.2 we calculate the entropy and information of some useful Matrix-variate distributions.

In Chapter 6 two tractable multivariate models are formulated. Section 6.1 introduces a general recursive model analogous to the fully recursive econometric models (Zellner, 1971). In Section 6.2 an example using artificial data illustrates an application to nested missing observations. In Section 6.3 we make use of the discount concept (Ameen and Harrison, 1983; Harrison and West, 1986) in order to simulate a DLMR with a dynamic scale variance. In Section 6.4 we compare the performance of a DWMR, with a dynamic scale variance matrix, to its static counterpart using exchange rate data. In addition, we employed PIE's for estimating the principal components of the scale variance matrix. This provides an insight into exchange rate dynamics. The use of simulation in Bayesian statistics is discussed in Appendix A6.1. This material is employed in the example of Section 6.2. A result concerning the distribution of swept matrices is provided in Appendix A6.2.

Chapter 7 is devoted to modelling aspects. The setting of vague priors for the DLMR model is briefly discussed in Section 7.1. The use of the multivariate logarithmic and logarithmic ratio transformations in the context of DWMR modelling is the topic of Section 7.2. Section 7.3 concerns alternative but equivalent reformulations of the DLMR model which expand its field of applications. The models considered are: perfect observations, colored observation error, colored evolution noise, correlated noise

and error, fixed-lag smoothing and prediction, differencing series and transfer response functions.

Finally, Chapter 8 consists of a general discussion and possible topics for further research.

The thesis is written in an informal style; the results presented are justified rather than rigorously proved.

1.5 NOTATION.

Throughout the thesis the following notation is employed unless otherwise specified. Scalars are denoted by lowercase letters (v, w, \dots), column vectors by underlined lowercase letters ($\underline{x}, \underline{y}, \dots$) and matrices by uppercase letters (X, Y, \dots) and unknown parameters by Greek letters (Θ, Σ, \dots). The factorial symbol ! is used in a extended sense, i.e. $x!$ denotes the gamma function evaluated at $x + 1$, and the symbol Γ is reserved to denote the gamma distribution. The digamma function - the derivative of the (natural) logarithm of the gamma function - is denoted by $\delta(x)$.

Expressions like $\alpha^{-1}A$ are often denoted by $\frac{A}{\alpha}$. Conformable matrices are partitioned preserving the conformability within the submatrices, e.g. in the expression

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix},$$

the products $AE, BG, AF, BH, CE, DG, CF$ and DH are implicitly assumed to be well defined. Similarly, in the expression

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} + \begin{bmatrix} E & F \\ G & H \end{bmatrix},$$

the sums $A + E, B + F, C + G$ and $D + H$ are assumed to be well defined, etc. In addition, we use the diag operator for denoting block diagonal matrices, e.g. $\text{diag}(A, B, C) = \begin{bmatrix} A & O & O \\ O & B & O \\ O & O & C \end{bmatrix}$.

In order to simplify the notation common conditional information is often omitted as in the specification of the dynamic model in Sections 1.2-3, e.g. the model (1.2) is implicitly assumed to be

$$\begin{aligned} p(\theta_t | H_{t-1}) &= \int_{\theta_{t-1}} p(\theta_t | \theta_{t-1}, H_{t-1}) p(\theta_{t-1} | H_{t-1}) d\theta_{t-1}, \\ p(y_t | H_{t-1}) &= \int_{\theta_t} p(y_t | \theta_t, H_{t-1}) p(\theta_t | H_{t-1}) d\theta_t, \quad \text{and} \\ p(\theta_t | y_t, H_{t-1}) &= \frac{p(y_t | \theta_t, H_{t-1}) p(\theta_t | H_{t-1})}{p(y_t | H_{t-1})}, \end{aligned}$$

where H_{t-1} stands for the relevant information known up to time $t - 1$. The random variables on which operators such as mean and variance depend are included in the notation as subscripts, e.g.

$$E_f(x) = \int_x f(x) p(x|\theta) dx.$$

Our notation and terminology for the models described in the following chapters is a compromise between the conventional notation and terminology in Bayesian forecasting and econometric literature. As any compromise, this has advantages and disadvantages depending on one's point of view. Further notation is introduced when needed, in particular in Appendices A3.1, A3.2 and A4.1.

CHAPTER 2

THE DYNAMIC LINEAR MODEL

Since the introduction of the Dynamic Linear Model (DLM) by Harrison and Stevens (1976), Bayesian forecasters have made use of an appealing dynamic model which can deal with multivariate time series. The DLM offers many facilities: use of prior information to start up the system, construction of complex models from simple components with the aid of the superposition principle, intervention analysis, discrimination between rival models, simple sequential updating recursions, joint forecast distributions, time invariant interpretation of the system parameters, and so on.

In this chapter the DLM is reviewed and a basis is provided for the next chapter where an extension of the DLM, the Dynamic Linear Matrix Regression Model (DLMR), is developed. The DLM description is given in Section 2.1, the updating recursions are provided in Section 2.2, time invariant interpretation of the parameters and model building from simple components are the topics in Section 2.3, Multivariate DLM's are entertained in Section 2.4, and the problem of specification of the evolution variance and observational variance with emphasis on the multivariate case is discussed in Section 2.5.

2.1 MODEL FORMULATION.

The assumptions of the DLM of Harrison and Stevens (1976) for a multivariate time series \underline{y}_t are:

Observation Equation:

$$\underline{y}_t = \underline{X}_t \underline{\theta}_t + \underline{\epsilon}_t, \quad \underline{\epsilon}_t \sim N(\underline{0}, V_t). \quad (2.1a)$$

Evolution Equation:

$$\underline{\theta}_t = \underline{G}_t \underline{\theta}_{t-1} + \underline{f}_t, \quad \underline{f}_t \sim N(\underline{0}, W_t). \quad (2.1b)$$

Prior Information:

$$\underline{\theta}_{t-1} \sim N(\underline{m}_{t-1}, C_{t-1}). \quad (2.1c)$$

Where,

- t is the time index ($t = 1, 2, 3, \dots$),
- \underline{y}_t is a $(r \times 1)$ vector of observations made at time t ,
- \underline{X}_t is a $(r \times p)$ matrix of independent variables,
- $\underline{\theta}_t$ is a $(p \times 1)$ unknown vector of system (regression) parameters,
- $\underline{\epsilon}_t$ is a $(r \times 1)$ observation error vector,
- V_t is a $(r \times r)$ variance matrix associated with $\underline{\epsilon}_t$,
- \underline{G}_t is a $(p \times p)$ evolution (trend) matrix,
- \underline{f}_t is a $(p \times 1)$ evolution noise vector,
- W_t is a $(p \times p)$ variance matrix associated with \underline{f}_t .

As usual, $N(\underline{m}, C)$ denotes the multivariate normal distribution with mean \underline{m} and variance C .

The equation (2.1a) defines the distribution of the observations \underline{y}_t given the system parameters. The system dynamics are determined by (2.1b) which specifies the distribution of the regression coefficients at time t in terms of the previous regression coefficients. The system prior information (2.1c) represents

the distribution of the regression coefficients at time $t - 1$. The distributions appearing in (2.1) are implicitly assumed conditional on the relevant information available at time $t - 1$ including (if any) previous observations $\underline{y}_{t-1}, \underline{y}_{t-2}, \dots$. It is further assumed that ε_t, f_t and θ_{t-1} are independent, with ε_t, f_t independent over time.

The following terminology is employed. When X_t, G_t, V_t and W_t are not dependent on time then the DLM is called a constant DLM. If in addition $W = 0$, then a DLM is referred to as a noise-free constant DLM.

2.2 UPDATE COMPUTATIONS.

When a new observation is obtained we have to revise our beliefs about the parameters (and implicitly our beliefs about the forthcoming observations). It is sufficient to describe this updating process for just one observation, since it can be performed over and over again when more observations become available, in a sequential fashion which is characteristic of the Bayesian approach.

The updating recursions for the DLM are:

Evolution:

$$\underline{\theta}_t \sim N(\underline{m}_t^*, C_t^*), \quad \text{where } C_t^* = W_t + G_t C_{t-1} G_t' \quad \text{and } \underline{m}_t^* = G_t \underline{m}_{t-1}. \quad (2.2a)$$

Prediction:

$$\underline{y}_t \sim N(\hat{y}_t, \check{Y}_t), \quad \text{where } \check{Y}_t = V_t + X_t C_t^* X_t' \quad \text{and } \hat{y}_t = X_t \underline{m}_t^*. \quad (2.2b)$$

Posterior:

$$\begin{aligned} \underline{\theta}_t | \underline{y}_t &\sim N(\underline{m}_t, C_t), \quad \text{where } C_t = C_t^* - A_t \check{Y}_t A_t', \quad \underline{m}_t = \underline{m}_t^* + A_t \hat{\varepsilon}_t, \\ &\quad \hat{\varepsilon}_t = \underline{y}_t - \hat{y}_t \quad \text{and } A_t = C_t^* X_t' \check{Y}_t^{-1}. \end{aligned} \quad (2.2c)$$

Equation (2.2a) provides the distribution of the regression coefficients at time t conditional to the information available at time $t - 1$. The one-step ahead predictive distribution of \underline{y}_t made at time $t - 1$ is given by (2.2b). The regression coefficients distribution at time t updated with the additional information is provided by (2.2c) completing the system cycle.

The derivation of (2.2) is a simple exercise of multivariate normal theory, this is not done here, but in the next chapter the updating recursions for the DLMR, which contains the DLM, are obtained. The system (2.1) is in probabilistic terms equivalent to the Bayesian formulation of the Kalman (1963) state-space model found, for example, in Aoki (1967) and Maybeck (1979). The counterpart of (2.2) is known as the Kalman filter in engineering literature.

The recursions (2.2) provide a very convenient way of updating the system. The statistic \underline{m}_t (together with C_t) is sufficient for $\underline{\theta}_t$ at time t and summarizes all past information. Thus, for forecasting purposes, it is equivalent to the whole previous history of the system. The implementation of (2.2) may be done in a straightforward manner with the possible exception of (2.2c), which requires, in the multivariate case, the computation of the matrix inverse of \check{Y}_t . The square-root algorithms found in the

Kalman filter literature, see for example Maybeck (1979) and Anderson and Moore (1979), are useful for dealing with systems requiring high precision. Perhaps the easiest implementation of (2.2) is by means of the Efroymson (1960) sweep operator. These implementation aspects are covered in Chapter 4.

2.3 STANDARD MODELS.

In state-space modelling the aim is to make inferences about the past, present and future (smoothing, filtering and prediction) of the system state $\underline{\theta}_t$, which usually has a physical meaning. In contrast, Bayesian forecasting is concerned, primarily, with the distribution of the future observations based on the available information. The vector parameter $\underline{\theta}_t$ represents time varying regression coefficients which, in the standard models, have a time invariant interpretation and play a very important role in understanding and operating the system.

2.3.1 Superposition Principle.

The superposition principle is a simple, but powerful statement, which tells us that a linear combination of DLM's is itself a DLM. In particular, it means that the trend of the series \underline{y}_t given by $X_t \underline{\theta}_t$ and $\underline{\theta}_t = G_t \underline{\theta}_{t-1} + \underline{w}_t$ can be constructed as the sum of simple trends, i.e. two trends

$$X_{it} \underline{\theta}_{it}, \quad \underline{\theta}_{it} = G_{it} \underline{\theta}_{i(t-1)} + \underline{w}_{it}, \quad \underline{w}_{it} \sim N(0, W_{it}), \quad i = 1, 2$$

can be combined into

$$X_t \underline{\theta}_t, \quad \underline{\theta}_t = G_t \underline{\theta}_{t-1} + \underline{w}_t, \quad \underline{w}_t \sim N(0, W_t)$$

where

$$\begin{aligned} X_t \underline{\theta}_t &= [X_{1t}, X_{2t}] \begin{bmatrix} \underline{\theta}_{1t} \\ \underline{\theta}_{2t} \end{bmatrix} = X_{1t} \underline{\theta}_{1t} + X_{2t} \underline{\theta}_{2t}, \\ G_t &= \begin{bmatrix} G_{11} & O \\ O & G_{22} \end{bmatrix}_t, \quad \text{and } W_t = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}_t. \end{aligned}$$

The generalization for several trends is evident.

This facility is very useful since it provides a method for building up complex models from simple components, and it is employed in one form or another in many models appearing in the following chapters.

2.3.2 The Polynomial Trend.

A univariate DLM with polynomial trend of degree $p - 1$ can be constructed by taking $X_t = X = [1, 0, \dots, 0]$ and $G_t = G$ a right triangular matrix such that the non-zero triangle is precisely the Pascal triangle (from left to right), i.e.

$$G_{ij} = \binom{j-1}{i-1} (i \leq j), \quad i, j = 1, \dots, p, \quad (2.3)$$

where $(i \leq j)$ means 1 if $i \leq j$ and 0 otherwise. A key property of G , shown in Appendix A2, is the following:

$$G_{ij}^* = \binom{j-1}{i-1} (i \leq j) s^{j-i}, \quad i, j = 1, \dots, p. \quad (2.4)$$

Notice that the elements of the first row of G^s are the powers of s , i.e. $1, s, \dots, s^{p-1}$.

For convenience we consider first the noise-free DLM i.e.

$$\underline{\theta}_t = G\underline{\theta}_{t-1}. \quad (2.5)$$

In this case, from (2.4) and (2.5) it is clear that the trend is given by

$$X\underline{\theta}_{t+s} = \dots = [1, (s-1), \dots, (s-1)^{p-1}] \underline{\theta}_{t+1} = [1, s, \dots, s^{p-1}] \underline{\theta}_t \quad (2.6)$$

Therefore, the trend has a polynomial form and the parameter $\underline{\theta}_t$ represents the polynomial coefficients relative to the system of coordinates with origin at $(0, t)$.

If the noise-free assumption is dropped, then the parameters have random disturbances but the previous interpretation remains valid by substituting the parameters with their expectations, i.e. the forecast function (the mean of y_{t+s} conditional on the information available at time t) is given by,

$$F_t(s) = X\underline{m}_{t+s} = \dots = [1, s, \dots, s^{p-1}] \underline{m}_t. \quad (2.7)$$

It should be noticed that other representations are possible, for instance, taking

$$X = [1, 0, \dots, 0] \quad \text{and} \quad G = I + \begin{bmatrix} 0 & I \\ 0 & 0' \end{bmatrix}$$

gives the Jordan-canonical form based on the standarized factorial polynomials $\binom{s}{0}, \binom{s}{1}, \dots, \binom{s}{p-1}$ rather than on the powers of s .

2.3.3 The Harmonic Trend.

Proceeding as before, a univariate DLM with simple harmonic trend of period $2\pi/\omega$ can be constructed by taking

$$X_t = X = [1, 0] \quad \text{and} \quad G_t = G = \begin{bmatrix} \cos \omega & \sin \omega \\ -\sin \omega & \cos \omega \end{bmatrix}.$$

The counterpart of (2.4) is

$$G^s = \begin{bmatrix} \cos \omega s & \sin \omega s \\ -\sin \omega s & \cos \omega s \end{bmatrix}, \quad (2.8)$$

which may be derived easily using the well-known trigonometric formulas for the addition of angles.

Again, we first consider the trend of the noise-free DLM. It follows from (2.5) and (2.8) that the trend is,

$$X\underline{\theta}_{t+s} = \dots = [\cos \omega(s-1), \sin \omega(s-1)] \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{t+1} = [\cos \omega s, \sin \omega s] \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_t. \quad (2.9)$$

Thus, the trend consists in a single harmonic and θ_{1t}, θ_{2t} are the coefficients associated with the cosine and sine components relative to the system of coordinates with origin at $(0, t)$.

As before, this interpretation remains valid when the noise-free assumption is dropped, if the parameters are substituted by their expectations, and the forecast function is,

$$F_t(s) = [1, 0] \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}_{t+s} = \dots = [\cos \omega s, \sin \omega s] \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}_t. \quad (2.10)$$

Complex harmonic trends with several harmonics are achieved easily by means of the superposition principle.

2.3.4 Damped Trends.

A damped version of a trend given by $X_t = X$ and $G_t = G$ is constructed by taking λG as the new trend matrix instead of G , where λ is a suitable scalar typically $0 < \lambda < 1$.

The mixtures of polynomial, harmonic and damped trends constructed with the aid of the superposition principle provide a very rich stock of models, certainly wide enough for the purposes of this thesis.

2.4 MULTIVARIATE MODELS.

All standard models discussed so far are univariate models. However, one version of the superposition principle provides a method for building-up multivariate DLM's from univariate DLM's.

Two multivariate DLM models for \underline{y}_{1t} and \underline{y}_{2t} ,

$$\underline{y}_{it} = X_i \theta_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(\underline{0}, V_{itt}), \quad i = 1, 2, \quad (2.11a)$$

$$\theta_{it} = G_{it} \theta_{i(t-1)} + \underline{w}_{it}, \quad \underline{w}_{it} \sim N(\underline{0}, W_{itt}), \quad i = 1, 2, \quad (2.11b)$$

can be combined into a single DLM for $\begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \end{bmatrix}_t$,

$$\begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \end{bmatrix}_t = \begin{bmatrix} X_{11} & O \\ O & X_{22} \end{bmatrix}_t \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_t + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}_t, \quad \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}_t \sim N\left(\underline{0}, \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}_t\right), \quad (2.12a)$$

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_t = \begin{bmatrix} G_{11} & O \\ O & G_{22} \end{bmatrix}_t \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{t-1} + \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}_t, \quad \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}_t \sim N\left(\underline{0}, \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}_t\right), \quad (2.12b)$$

The generalization for several components is easily appreciated. For instance, a trivariate DLM, with a linear, simple harmonic and damped constant trend for the first, second and third dependent variables respectively, has a regressor matrix

$$\text{diag}([1, 0], [1, 0], 1) \quad \text{and a trend matrix } \text{diag}\left(\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} \cos \omega & \sin \omega \\ -\sin \omega & \cos \omega \end{bmatrix}, \lambda\right).$$

A simple but very useful DLM for time series in which all the variables have a similar trend determined by X_t and G_t , has a regressor matrix $\text{diag}(X, \dots, X)$ and a trend matrix $\text{diag}(G, \dots, G)$. The importance of this class of models becomes apparent in the following chapters.

2.5 SPECIFICATION OF THE VARIANCES.

In order to implement a DLM, practitioners face a major obstacle: the setting of two system variances W_t and V_t . The first difficulty has been overcome partially by Ameen and Harrison (1984) through the discount concept which substitutes the evolution variance matrix W_t by a set of discount factors and offers a conceptually simple alternative model. Nevertheless, the specification of the observational variance V_t for multivariate DLM's remains a major problem for practitioners.

2.5.1 The Discount Weighted Version.

The DLM version, based on the discount concept, merely substitutes the first equation in (2.2a) by,

$$C_t^* = (\text{diag } \beta)^{-\frac{1}{2}} G_t C_{t-1} G_t' (\text{diag } \beta)^{-\frac{1}{2}} \quad (2.13)$$

where β is a set of discount factors, i.e. each element of β lies between 0 and 1.

The advantage of using the discount version over the standard version is clear; the specification of a rather cumbersome variance W_t is traded for a set of discount factors for which many practitioners have a natural feeling. Often a discount model can be interpreted as a DLM; however, the equivalence between a discount DLM and a standard DLM is not always guaranteed, i.e. the matrix $(\text{diag } \beta)^{-\frac{1}{2}} G_t C_{t-1} G_t' (\text{diag } \beta)^{-\frac{1}{2}} - G_t C_{t-1} G_t'$ fails sometimes to be a proper variance. Furthermore, the equivalence, for an arbitrary discount set β , is guaranteed if and only if $G_t C_{t-1} G_t'$ is diagonal.

The discount version may be modified, in order to assure equivalence by adopting the following rule instead of (2.13),

$$C_{ijt}^* = \begin{cases} \beta_i^{-1} B_{ijt}, & \text{if } \beta_i = \beta_j, \\ B_{ijt}, & \text{otherwise} \end{cases}, \quad (2.14)$$

where $B_t = G_t C_{t-1} G_t'$. A parsimonious form for independent components recommended by West and Harrison (1986) is to take the same discount factor within elements of each component. This procedure is referred to as discount by blocks. In some circumstances it is useful to combine a discount procedure followed by the addition of a noise variance.

2.5.2 On-line Variance Learning.

A natural alternative for avoiding the specification of the observational variance is to assume that it is unknown. This approach, in the univariate case, replaces the model (2.1) by

Observation Equation:

$$y_t = \underline{x}_t' \underline{\theta}_t + e_t, \quad e_t \sim N(0, \sigma^2 v_t) \quad (2.15a)$$

Evolution Equation:

$$\underline{\theta}_t = G_t \underline{\theta}_{t-1} + f_t, \quad f_t \sim N(0, \sigma^2 W_t) \quad (2.15b)$$

Prior Information:

$$\underline{\theta}_{t-1} \sim N(\underline{m}_{t-1}, \sigma^2 C_{t-1}) \quad \text{and} \quad \sigma^2 \sim \Gamma^{-1}(\frac{1}{2} d_{t-1}, \frac{1}{2} s_{t-1}). \quad (2.15c)$$

Where $\sigma^2 \sim \Gamma^{-1}(\frac{1}{2} d, \frac{1}{2} s)$ means that σ^{-2} has a gamma distribution with mean $\frac{1}{2}s$ and variance $\frac{1}{2}s^2$. It is implicitly assumed that the distributions in (2.15a), (2.15b) and in the left-hand side of (2.15c) are conditional on σ^2 and also, of course, on the relevant information available at time $t-1$. Henceforth we denote the joint distribution induced by (2.15c) as $\left[\begin{array}{c} \underline{\theta}_{t-1} \\ \sigma^2 \end{array} \right] \sim N\Gamma^{-1}(\underline{m}_{t-1}, C_{t-1}, \frac{1}{2} d_{t-1}, \frac{1}{2} s_{t-1})$ and the marginal distribution of $\underline{\theta}_{t-1}$ as $\underline{\theta}_{t-1} \sim t(\underline{m}_{t-1}, s_{t-1} C_{t-1}, d_{t-1})$.

Using the recursions (2.2a) and (2.2c) for $\theta_t|\sigma^2$ and (2.2b), together with a direct application of Bayes' formula for σ^2 , the following updating recursions are readily obtained,

Evolution:

$$\left[\frac{\theta_t}{\sigma^2} \right] \sim N \Gamma^{-1}(\underline{m}_t^*, C_t^*, \frac{1}{2}d_{t-1}, \frac{1}{2}s_{t-1}) \quad (2.16a)$$

Prediction:

$$y_t \sim t(\hat{y}_t, s_{t-1}\check{y}_t, d_{t-1}) \quad (2.16b)$$

Posterior:

$$\left[\frac{\theta_t}{\sigma^2} \right] | y_t \sim N \Gamma^{-1}(\underline{m}_t, C_t, \frac{1}{2}d_t, \frac{1}{2}s_t) \quad (2.16a)$$

where $\underline{m}_t^*, C_t^*, \hat{y}_t, \check{y}_t, \hat{e}_t$ are as in (2.2) and $d_t = d_{t-1} + 1, s_t = s_{t-1} + \hat{e}_t^2/\check{y}_t$.

Clearly, for $\sigma^2 = 1$ the model (2.15) becomes a conventional univariate DLM. Therefore, it is an extension of the DLM model and describes essentially the model employed by Smith and West (1983), a non-Bayesian formulation can be found in Harvey (1984). It is important to notice that, for $v_t = 1$, σ^2 is not only the unknown observational variance but it also appears as a scalar factor in the evolution variance.

The above procedure can be easily extended to the multivariate case when the observational variance is assumed known except by an unknown scale factor, and it is found in West (1982). Unfortunately, any attempt to leave the observational variance unknown as a whole in the multivariate case faces the problem of intractability, the reason being that the generality of the likelihood rather than the dynamic structure of the model. For instance, the static DLM leads to an intractable analysis as is shown in the next chapter. Hence, the specification of the observational variance in the multivariate case remains a major obstacle. Furthermore, one of the main objectives of an analysis may well be to learn about the observational variance.

APPENDIX A2.1.

THE POLYNOMIAL TRANSITION MATRIX.

The representation of a polynomial trend in terms of powers of s , although natural, is not common in Bayesian forecasting. For this reason we give a derivation of (2.4).

The formula (2.4) may be verified by means of double induction, on p and s , as follows. Let

$$G^s = \begin{bmatrix} G_{11}^{(s)} & G_{12}^{(s)} \\ 0 & 1 \end{bmatrix},$$

then it is enough to show that

$$\begin{bmatrix} G_{11}^{(1)} & G_{12}^{(1)} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} G_{11}^{(s-1)} & G_{12}^{(s-1)} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} G_{11}^{(s)} & G_{12}^{(s)} \\ 0 & 1 \end{bmatrix},$$

but $G_{11}^{(s)} G_{11}^{(s-1)} = G_{11}^{(s)}$ by the induction hypothesis on p , and using the binomial theorem it is clear that

$$\begin{aligned} G_{11}^{(1)} G_{12}^{(s-1)} + G_{12}^{(1)} &= \left[\binom{j-1}{i-1} (i \leq j) \right] \left[\binom{p-1}{j-1} (s-1)^{p-j} \right] + \left[\binom{p-1}{i-1} \right] \\ &= \left[\sum_{j=i}^{p-1} \binom{j-1}{i-1} \binom{p-1}{j-1} (s-1)^{p-j} + \binom{p-1}{i-1} \right] \\ &= \left[\sum_{j=i}^{p-1} \binom{p-1}{i-1} \binom{p-i}{p-j} (s-1)^{p-j} + \binom{p-1}{i-1} \right] \\ &= \left[\binom{p-1}{i-1} \sum_{j=0}^{p-i} \binom{p-i}{j} (s-1)^{p-i-j} \right] = G_{12}^{(s)}, \end{aligned}$$

since $\binom{j-1}{i-1} \binom{p-1}{j-1} = \binom{p-1}{i-1} \binom{p-i}{p-j}$.

CHAPTER 3

DYNAMIC LINEAR MATRIX-VARIATE REGRESSION

This chapter is concerned with a dynamic linear model proposed in Quintana (1985), which extends the DLM by considering matrices of observations, instead of vectors, and introducing a new system scale variance matrix. In doing so, the dynamic structure of the DLM is transferred to the new model which includes the standard multivariate regression model, with the observational variance unknown, as a special, static case. This matrix-variate model is referred to as the Dynamic Linear Matrix-variate Regression (DLMR).

In 3.1 the intractability of the static DLM with an unknown observational variance is shown, the DLMR is suggested and the description of the DLMR is given. Its updating recurrences are derived in 3.2. Static and dynamic models contained in the DLMR, a vector representation, a discount version and model building facilities inherited from the DLM are entertained in 3.3. Finally, multi-process models, reparameterization and system decoupling are discussed in Sections 4, 5 and 6 respectively.

Henceforth several results of the vec operator, Kronecker product \otimes , and matrix-variate distribution theory are invoked freely. Definitions and properties of the Kronecker product and the vec operator are provided in Appendix A3.1. The essential singular and non-singular matrix variate distribution theory is developed in Appendix A3.2.

3.1 MODEL FORMULATION.

In this section we confront the difficulties associated with the DLM when the error variance is assumed unknown. Then, perceiving how the problem is overcome in the static case we formulate a tractable dynamic model.

3.1.1 Static and Dynamic Models.

In Subsection 2.5.2 it is claimed that the static DLM with the constant variance V unknown leads to an intractable analysis. This may be shown as follows. This static model is the DLM (2.1) with $G_t = I$, $V_t = V$ and $W_t = O$ for all t , i.e. its observation equation is given by,

$$\underline{y}_t = X_t \underline{\theta} + \underline{\epsilon}_t, \quad t = 1, \dots, n, \quad \underline{\epsilon}_t \sim N(O, V). \quad (3.1a)$$

It can be rewritten in the matrix form

$$Y = X\Theta + E, \quad E \sim N(O, V, I), \quad (3.1b)$$

where $Y = [\underline{y}_1, \dots, \underline{y}_n]$ is a $(r \times n)$ matrix, $X = [X_1, \dots, X_n]$ is a $(r \times (pn))$ matrix, $\Theta = I \otimes \underline{\theta}$ is a $((pn) \times n)$ matrix, I is a $(n \times n)$ matrix, and $E = [\underline{\epsilon}_1, \dots, \underline{\epsilon}_n]$ is a $(r \times n)$ matrix. Therefore the likelihood is such that (see A3.2.10)

$$p(Y|\underline{\theta}, V) \propto |V|^{-\frac{r}{2}} \exp(-\frac{1}{2} \text{tr}(EE'V^{-1})). \quad (3.2)$$

Problems arise because there is not in the general case a tractable sufficient statistic; see Section 1.3. This difficulty is essentially the same as that found in related static linear models such as the general linear model with common regression coefficients (Box and Tiao, 1973, p. 501-502) and the seemingly unrelated regression (Zellner, 1971, p. 240-243). Furthermore, let us assume that the variances in the dynamic model (2.1) are constant ($V_t = V$ and $W_t = W$ for all t) but unknown with independent inverted-Wishart distributions. In this case it can be shown that the posterior distribution of $\underline{\theta}_t$ is an intractable multivariate poly-t (Broemeling, 1985, p. 286-290).

Hence, it is clear that extra assumptions are necessary in order to obtain a tractable procedure for on-line variance learning. The artifice for the static case is well-known; the standard multivariate regression is considered as an alternative to a more general model such as the generalized weighted multivariate regression model. Both models can be embedded in the matrix model,

$$Y = X\Theta + E, \quad E \sim N(O, V, \Sigma), \quad (3.3a)$$

with the conjugate joint matrix-normal inverted-Wishart prior for Θ, Σ given by

$$\Theta \sim N(M, C, \Sigma) \quad \text{and} \quad (3.3b)$$

$$\Sigma \sim W^{-1}(S, d). \quad (3.3c)$$

It is further assumed that Θ and E are independent given Σ .

Taking $V = I$ we have the standard multivariate regression with the observational variance Σ unknown. When Y is a column, and $X = \text{diag}(X_1, \dots, X_n)$, we obtain the weighted generalized multivariate regression where the observational variance $V\Sigma$ is known except for an unknown scalar factor Σ ; compare with Press (1982, p. 245-248).

To see how the discussion of the static case brings some insight to the dynamic case, we have to realize that a dynamic model can be thought of as a static one by considering the evolution and the prior as a joint prior, e.g. the DLM (2.15) can be seen at time t as,

$$\underline{y} = X\underline{\theta} + \underline{\varepsilon}, \quad \underline{\varepsilon} \sim N(\underline{0}, \sigma^2 V), \quad (3.4a)$$

$$\underline{\theta} \sim N(\underline{m}, \sigma^2 C), \quad (3.4b)$$

where

$$\begin{aligned} \underline{y} &= \underline{y}_t, & X &= [O, X_t], & \underline{\theta} &= \begin{bmatrix} \underline{\theta}_{t-1} \\ \underline{\theta}_t \end{bmatrix}, & \underline{\varepsilon} &= \underline{\varepsilon}_t, & V &= V_t, \\ \underline{m} &= \begin{bmatrix} I & O \\ G_t & I \end{bmatrix} \begin{bmatrix} \underline{m}_{t-1} \\ \underline{0} \end{bmatrix} & \text{and } C &= \begin{bmatrix} I & O \\ G_t & I \end{bmatrix} \begin{bmatrix} C_{t-1} & O \\ O & W_t \end{bmatrix} \begin{bmatrix} I & O \\ G_t & I \end{bmatrix}' \end{aligned}$$

since

$$\begin{bmatrix} \underline{\theta}_t \\ \underline{\theta}_{t-1} \end{bmatrix} = \begin{bmatrix} I & O \\ G_t & I \end{bmatrix} \begin{bmatrix} \underline{\theta}_{t-1} \\ \underline{f}_t \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \underline{\theta}_{t-1} \\ \underline{f}_t \end{bmatrix} \sim N \left(\begin{bmatrix} \underline{m}_t \\ \underline{0} \end{bmatrix}, \begin{bmatrix} C_{t-1} & O \\ O & W_t \end{bmatrix} \right).$$

The representation (3.4) of the DLM (2.15) suggests that we consider the model (3.3) at time t where

$$Y = Y_t, \quad X = [O, X_t], \quad \Theta = \begin{bmatrix} \Theta_{t-1} \\ \Theta_t \end{bmatrix}, \quad E = E_t, \quad V = V_t,$$

$$M = \begin{bmatrix} M_{t-1} \\ G_t M_{t-1} \end{bmatrix}, \quad C = \begin{bmatrix} C_{t-1} & C_{t-1} G'_t \\ G_t C_{t-1} & W_t + G_t C_{t-1} G'_t \end{bmatrix}, \quad S = S_{t-1} \quad \text{and } d = d_{t-1},$$

i.e.

Observation Equation:

$$Y_t = X_t \Theta_t + E_t, \quad E_t \sim N(0, V_t, \Sigma). \quad (3.5a)$$

Evolution Equation:

$$\Theta_t = G_t \Theta_{t-1} + F_t, \quad F_t \sim N(0, W_t, \Sigma). \quad (3.5b)$$

Prior Information:

$$\begin{bmatrix} \Theta_{t-1} \\ \Sigma \end{bmatrix} \sim NW^{-1}(M_{t-1}, C_{t-1}, S_{t-1}, d_{t-1}). \quad (3.5c)$$

Where E_t, F_t and Θ_{t-1} are independent (with E_t, F_t independent over time) given Σ ; see A3.2.8.

3.1.2 Model Description.

Throughout the thesis the model (3.5) is referred to as the Dynamic Linear Matrix-variate Regression (DLMR) and the following notation is used:

- Y_t is a $(r \times q)$ matrix of observations made at time t ,
- X_t is a $(r \times p)$ matrix of independent variables,
- Θ_t is an unknown $(p \times q)$ matrix of system (regression) parameters
- E_t is a $(r \times q)$ observation error matrix,
- V_t is a $(r \times r)$ variance matrix associated with E_t ,
- G_t is a $(p \times p)$ evolution (trend) matrix,
- F_t is a $(p \times q)$ evolution noise matrix,
- W_t is a $(p \times p)$ variance matrix associated with F_t , and
- Σ is a $(q \times q)$ system scale matrix.

The matrices X_t, V_t, G_t and W_t are assumed to be known before Y_t is observed. In general it is supposed that Σ is unknown with prior inverted Wishart distribution, at time $t-1$ (after Y_{t-1} is observed), given by

$$\Sigma \sim W^{-1}(S_{t-1}, d_{t-1}) \quad (3.6a)$$

in accordance with (3.5c). However, in some circumstances it is convenient to assume that Σ is given. Naturally in this latter case,

$$\Theta_{t-1} \sim N(M_{t-1}, C_{t-1}, \Sigma), \quad (3.6b)$$

since (3.5c) and (3.6) are equivalent by definition; see A3.2.17.

Several useful static and dynamic multivariate models are embedded in this model, a list is given in Subsection 3.3.1. The role of Σ as a matrix scale factor is analogous as that of σ^2 in model (2.15). This is discussed in Subsection 3.3.2.

3.2 UPDATE COMPUTATIONS.

Our goal is to derive and to analyse the asymptotic behaviour of the DLMR updating recursions analogous to the recurrences for the DLM. These resulting recursions enable us to perform the process of incorporating the information of the observations in the usual sequential fashion.

The most economical way of obtaining the sequential updating recursions for the DLMR is by borrowing the DLM results as in Quintana (1985). However, here we prefer to give a self-contained derivation based on the results shown in Appendix A3.2.

3.2.1 Updating Recurrences.

First let us derive, for convenience, the recursions analogous to the formulas (2.2). Assuming that Σ is given, the updating recursions for the model (3.5a), (3.5b) and (3.6b) are as follows.

Evolution:

$$\Theta_t \sim N(M_t^*, C_t^*, \Sigma), \quad \text{where } C_t^* = W_t + G_t C_{t-1} G_t' \quad \text{and } M_t^* = G_t M_{t-1}. \quad (3.7a)$$

Prediction:

$$Y_t \sim N(\hat{Y}_t, \check{Y}_t, \Sigma), \quad \text{where } \check{Y}_t = V_t + X_t C_t^* X_t' \quad \text{and } \hat{Y}_t = X_t M_t^*. \quad (3.7b)$$

Posterior:

$$\begin{aligned} \Theta_t | Y_t &\sim N(M_t, C_t, \Sigma), \quad \text{where } C_t = C_t^* - A_t \check{Y}_t A_t', \quad M_t = M_t^* + A_t \hat{E}_t, \\ & \quad \hat{E}_t = Y_t - \hat{Y}_t \quad \text{and } A_t = C_t^* X_t' \check{Y}_t^{-1}. \end{aligned} \quad (3.7c)$$

These recursions may be shown as follows. Using the independence of Θ_{t-1} and F_t together with (3.6b), (3.5b) and (A3.2.5) it is clear that Θ_t is distributed according to (3.7a) since

$$\begin{bmatrix} \Theta_{t-1} \\ F_t \end{bmatrix} \sim N\left(\begin{bmatrix} M_{t-1} \\ O \end{bmatrix}, \begin{bmatrix} C_{t-1} & O \\ O & W_t \end{bmatrix}, \Sigma\right) \quad \text{and } \Theta_t = [G_t, I] \begin{bmatrix} \Theta_{t-1} \\ F_t \end{bmatrix}.$$

Following a similar argument, (3.7a) and (3.5a) imply that

$$\begin{bmatrix} \Theta_t \\ Y_t \end{bmatrix} \sim \begin{bmatrix} I & O \\ X_t & I \end{bmatrix} \begin{bmatrix} \Theta_t \\ E_t \end{bmatrix} \quad \text{and } \begin{bmatrix} \Theta_t \\ E_t \end{bmatrix} \sim N\left(\begin{bmatrix} M_t^* \\ O \end{bmatrix}, \begin{bmatrix} C_t^* & O \\ O & V_t \end{bmatrix}, \Sigma\right),$$

i.e.

$$\begin{bmatrix} \Theta_t \\ Y_t \end{bmatrix} \sim N\left(\begin{bmatrix} M_t^* \\ X_t M_t^* \end{bmatrix}, \begin{bmatrix} C_t^* & C_t^* X_t' \\ X_t C_t^* & X_t C_t^* X_t' + V_t \end{bmatrix}, \Sigma\right). \quad (3.8)$$

Thus, the prediction and posterior equations (3.7b) and (3.7c) are simply the relevant marginal and conditional distributions of (3.8) according to (A3.2.8).

The recurrences (3.7) provide a recursive algorithm for updating the system since M_t and C_t summarize all present and past information at time t . It is of note that neither M_t nor C_t depend on Σ . So far, the recursions (3.10) are valid for any proper variances V_t, W_t, C_{t-1} and Σ , i.e. they may be singular. Only \check{Y}_t has been implicitly assumed non-singular, but even this requirement is abandoned in Section 4.2.

We now drop the assumption that Σ is known and instead it follows an inverted Wishart distribution at time $t - 1$ given by (3.7a). The conditional distribution of $\Sigma|Y_t$ is implicitly given by (3.7b) and (3.6a), since according to (A3.2.17),

$$\begin{bmatrix} Y_t \\ \Sigma \end{bmatrix} \sim \text{NW}^{-1}(\hat{Y}_t, \check{Y}_t, S_{t-1}, d_{t-1}),$$

and applying (A3.2.19) we have,

$$\Sigma|Y_t \sim W^{-1}(S_t, d_t), \quad (3.9)$$

where,

$$S_t = S_{t-1} + E_t' \check{Y}_t^{-1} E_t \quad \text{and} \quad d_t = d_{t-1} + 1.$$

Therefore, the recurrences (3.7) may be generalized for the model (3.5) as follows.

Evolution:

$$\begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix} \sim \text{NW}^{-1}(M_t^*, C_t^*, S_{t-1}, d_{t-1}). \quad (3.10a)$$

Prediction:

$$Y_t \sim T(\hat{Y}_t, \check{Y}_t, S_{t-1}, d_{t-1}). \quad (3.10b)$$

Posterior:

$$\begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix} | Y_t \sim \text{NW}^{-1}(M_t, C_t, S_t, d_t). \quad (3.10c)$$

Where $M_t^*, C_t^*, \check{Y}_t, \hat{Y}_t, M_t, C_t, \hat{E}_t$ and A_t are as in (3.7), and S_t, d_t are as in (3.9).

Similar comments as those following the derivation of recursions (3.7) apply to (3.10), in particular, (3.10) is valid even if S (Σ) is singular. The essential difference is that (3.9) provides an effective procedure for on-line learning about the system scale variance Σ . Long-term forecasting can be achieved by means of repetitive use of (3.10a), (3.10b) and replacing the recurrence in (3.10c) by $C_t = C_t^*$ and $M_t = M_t^*$ (provided, of course, that the driving parameters G_{t+s}, W_{t+s} and the observational parameters V_{t+s} and X_{t+s} are known for $s = 1, 2, \dots$).

3.2.2 Limiting Behaviour.

For simplicity we look first at the case for Σ known. The resemblance between (3.7) and (2.2) is evident, in fact (3.7) becomes (2.2) when $\Sigma = 1$. Moreover, the behaviour analysis of C_t and A_t can be reduced to that of the DLM because from (2.2) and (3.7) clearly M_t, C_t, A_t , etc., are computed exactly as if the columns of Y_t were driven by DLM's with common parameters X_t, G_t, V_t and W_t , regardless of the actual Σ .

A particular but very important case is the observable constant DLMR. This case is obtained when the associated DLM's, described in the previous paragraph, are observable constant DLM's, i.e. when X, G, V , and W are not time dependent and there is a vector \underline{h} such that,

$$\begin{bmatrix} \underline{h}' X \\ \underline{h}' X G \\ \vdots \\ \underline{h}' X G^{p-1} \end{bmatrix}$$

is a full rank matrix. In this case A_t, C_t, C_t^* and \check{Y}_t converge because the result holds for the DLM, for a simple proof of the latter see Harrison (1985). If the DLMR is unobservable, then only partial convergence can be assured, see again the reference above for details.

These results still hold when Σ is distributed according to (3.6a) instead of being known, since the behaviour of M_t, C_t, A_t , etc., is independent of the actual Σ as is mentioned in Subsection 3.2.1. However, in this case it is interesting to look at the behaviour of S_t and d_t as t increases. From (3.9) we obtain,

$$S_t = S_0 + \sum_{r=1}^t \hat{E}_r' \check{Y}_r^{-1} \hat{E}_r \quad \text{and } d_t = d_0 + t r \quad (3.11)$$

Therefore, in the limit, $\Sigma = \lim_{t \rightarrow \infty} \frac{S_t}{d_t - 2}$ in probability and (3.10) becomes (3.7).

These limiting results may be employed in order to reduce the update computations for a constant model after a period of time. Typically, the convergence is fast and the recursions (3.10) may be replaced by to $M_t^* = GM_{t-1}, \hat{Y}_t = XM_t^*$ and $M_t = M_t^* + A(Y_t - \hat{Y}_t)$, where A is the limiting value of A_t .

3.3 THE DLMR AS A DLM.

Modelling with the DLMR essentially can be reduced to modelling with the DLM because most of the DLM structure is inherited by the DLMR. In this section it is illustrated how this transference can be done. First, we look at some standard multivariate linear models which are embedded in the DLMR.

3.3.1 Models Contained.

The DLMR contains a wide variety of normal linear models which can be divided into two main categories: dynamic and static. By a static model we mean a DLMR which is not evolving at all i.e. $G_t = I$ and $W_t = O$ for all t . The particular settings are as follows.

Static Models:

(a) Standard Multivariate Regression. This model can be seen either sequentially as a DLMR with $r = 1$ and $V_t = 1$ for $t = 1, \dots, n$, or in its entirety as a DLMR with $r = n$ and $V_t = I$ at a fixed time $t = t_0$.

(b) Weighted Generalized Multivariate Regression. This model as defined in Press (1982) corresponds to a DLMR with $q = 1$ at a fixed time $t = t_0$. A classical interpretation of the hyperparameters M_t, C_t and S_t is given in the next chapter.

Dynamic Models:

(a) Dynamic Weighted Multivariate Regression (DWMR). The DWMR generalizes the Standard Multivariate Regression and corresponds to a DLMR with $r = 1$. Therefore, it is a very important case of the DLMR and it is discussed in Section 3.7.

(b) Multivariate Dynamic Linear Model. The DLM's correspond to DLMR's with $q = 1$ and therefore $\Sigma = \sigma^2$ is a scalar. For the Harrison and Stevens (1976) DLM, σ^2 is assumed to be known and

equal to one. For the multivariate extension of the DLM (2.15), σ^2 is assumed unknown and $\sigma^2 \sim \Gamma^{-1}(\frac{1}{2}d_{t-1}, \frac{1}{2}s_{t-1})$.

Thus, the DLMR can be thought of as a combination of extended DLM's and DWMR's. The columns of Y_t are being modelled marginally as extended DLM's with common parameters X_t, G_t, V_t and W_t whilst the rows are marginally modelled as DWMR's, where their parameters are the corresponding rows of X_t and diagonal elements V_t , and common evolution parameters G_t and W_t . These comments give full support to the discussion in the previous section. In accordance with the terminology of the static and dynamic models the DLMR is referred to as weighted or non-weighted depending on V_t (this must not be confused with the discount weighted method of Subsections 2.5.1 and 3.3.3).

The DLMR not only provides a theoretical framework for the models mentioned above, but also makes possible the implementation of a relatively simple program suitable for a common personal microcomputer which can handle all these models at once as is shown in the next chapter.

3.3.2 Vector Representation.

Readers already familiar with the DLM may find a vector representation of the DLMR more easily interpretable. Applying the vec operator on (3.5a), (3.5b) and (3.6b), and using properties listed in Appendices (A3.1) and (A3.2) we can readily rewrite the model (3.5) as,

Observation Equation:

$$\text{vec } Y_t = (I \otimes X_t) \text{vec } \Theta_t + \text{vec } E_t, \quad \text{vec } E_t \sim N(\underline{0}, \Sigma \otimes V_t). \quad (3.12a)$$

Evolution Equation:

$$\text{vec } \Theta_t = (I \otimes G_t) \text{vec } \Theta_{t-1} + \text{vec } F_t, \quad \text{vec } F_t \sim N(\underline{0}, \Sigma \otimes W_t). \quad (3.12b)$$

Prior Information:

$$\text{vec } \Theta_{t-1} | \Sigma \sim N(\text{vec } M_{t-1}, \Sigma \otimes C_{t-1}) \quad \text{and } \Sigma \sim W^{-1}(S_{t-1}, d_{t-1}). \quad (3.12c)$$

The representation (3.12) implies that, given Σ , the DLMR is a special case of the DLM (2.1), in which the driving parameters have a special structure given by the Kronecker direct product: Σ is a scale factor of the variances and I is a scale factor in the linear transformations. This shows the richness of the DLM and DLMR since, given Σ , each is a special case of the other!

The vector form of the DLMR is very useful in order to transfer the structure from the DLM to the DLMR. For instance, it is apparent from (3.12) that dropping the assumptions about normality, $\hat{\Theta}_t = M_t$ and $\hat{Y}_t = X_t \hat{\Theta}_t$ are the best linear estimators in the linear Bayesian sense (Hartigan, 1969), regardless of whether Σ is known or not. As mentioned before the DLMR updating recurrences (3.7), given Σ , may be derived using the DLM recurrences corresponding to (3.12) and switching back to the matrix representation (3.5a), (3.5b) and (3.6b), then the derivation for Σ unknown continues as before. In the following section we illustrate this procedure considering the discount version of the DLMR.

3.3.3 Discount Weighted Version.

The interpretation of the DLMR as a combination of extended DLM's and DWMR's given in Sub-section 3.3.1 suggests the use of the same discount factors $\underline{\beta}$ for each column of Θ_t . Thus, according to (2.13), (3.12) and (2.2a), the discount version substitutes,

$$(\Sigma \otimes W_t) + (I \otimes G_t)(\Sigma \otimes C)(I \otimes G_t)' = \Sigma \otimes (W_t + G_t C_{t-1} G_t') = \Sigma \otimes C_t^*,$$

by,

$$\begin{aligned} (\Sigma \otimes C_t)^* &= (I \otimes \text{diag } \underline{\beta})^{-\frac{1}{2}} (I \otimes G_t)(\Sigma \otimes C_{t-1})(I \otimes G_t)' (I \otimes \text{diag } \underline{\beta})^{-\frac{1}{2}} \\ &= \Sigma \otimes (\text{diag } \underline{\beta})^{-\frac{1}{2}} G_t C_{t-1} G_t' (\text{diag } \underline{\beta})^{-\frac{1}{2}} = \Sigma \otimes C_t^*, \end{aligned}$$

i.e. the first equation in (3.7a) is replaced by (2.13). For the reasons expressed in Section 2.5.1, the use of the rule of discount by blocks (2.14) is recommended instead of (2.13).

The analysis of the limiting behaviour can be reduced, in a similar manner as with the usual DLMR, to that of the DLM with a discount version found in Harrison (1985). For instance, an observable constant discount DLMR converges, etc.

3.3.4 Model Building.

Two aspects need to be considered when constructing a DLMR model: trend and variance structure. The standard trend models of Sections 2.3 and 2.4 can be built in a straight forward manner. It is only necessary to keep in mind that X_t and G_t determine the common trend form for the columns of Y_t . For instance, the setting given in Section 2.4,

$$X_t = \text{diag}([1, 0], [1, 0], 1) \quad \text{and} \quad G_t = \text{diag}\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} \cos \omega & \sin \omega \\ -\sin \omega & \cos \omega \end{bmatrix}, \lambda\right),$$

for a DLMR with $r = 3$, $p = 6$ and $q = 2$ means that the three bivariate rows of Y_t have a linear, a simple harmonic and a damped constant trend respectively.

Regarding the variance structure, V_t , W_t and C_t retain essentially the same meaning as with the DLM; V_t represents the inverse weight or precision of the present information, C_{t-1} measures the inverse weight of past information, and W_t controls the inverse weight of the link between past and present information.

In practice it is helpful to remember that V_t , W_t , C_{t-1} and the columns of Y_t steer C_t and M_t with the DLM recurrences (2.2). In particular modelling with a DWMR is essentially as easy (or difficult) as modelling with univariate DLM's.

Special models such as those with correlated error and noise, time correlated noise, transfer response functions, etc., are discussed in Chapter 7.

3.4 MULTI-PROCESS MODELS.

It is often the case that a modeller has in mind two or more possible models for a certain time series; to handle this, Harrison and Stevens (1971, 1976) introduced the multi-process models methodology for

discriminating between rival DLM's. Two situations are considered; multi-process models class I deals with fixed DLM's whilst in multi-process models class II the possible DLM's may change from one time interval to another following a Markovian scheme. The technical details of multiprocess models for the DLMR are given in this section.

3.4.1 Class I.

Suppose that we are considering a set of possible DLMR models $\mathcal{M}^{(i)}$ ($i = 1, 2, \dots, m$) for describing the time series Y_t ($t = 1, 2, \dots$). In addition, we believe that at least one of these models is adequate, but we do not know precisely which one is. Our problem is to discriminate between these rival models. Let $P_t(\mathcal{M}^{(i)})$ denote the probability that the time series follows the model $\mathcal{M}^{(i)}$ given all information available at time t but prior to observing Y_t , and $P_t(\mathcal{M}^{(i)}|Y_t)$ the revised probability given the additional information Y_t . Then, the two probabilities are related according to Bayes' theorem as follows,

$$P_t(\mathcal{M}^{(i)}|Y_t) \propto p(Y_t|\mathcal{M}^{(i)}) P_t(\mathcal{M}^{(i)}), \quad (3.13)$$

where $p(Y_t|\mathcal{M}^{(i)})$ denotes the predictive density for Y_t assuming the model $\mathcal{M}^{(i)}$, given all information available at time t .

The applicability of (3.13) depends on the existence of the predictive density for Y_t . But for the DLMR this is obtainable either from (3.7b) or (3.10b) depending on whether Σ is known or not. Therefore, the recurrence (3.13) provides a means for an effective on-line model discriminating procedure.

The possible DLMR models can be completely arbitrary; typically each $\mathcal{M}^{(i)}$ has its own associated parameters $X_t^{(i)}, G_t^{(i)}, V_t^{(i)}, W_t^{(i)}$ for $t = 1, 2, \dots$. The multi-process models class I offer a simple method for tuning constant DLMR's provided, of course, that the number of the rival models (m) remains manageable.

3.4.1 Class II.

Now we turn to the case in which a fixed DLMR may not adequately describe the time series. Instead it is assumed that the model may switch from one time interval to the next in a Markovian fashion. Let $\mathcal{M}_t^{(j)}$ ($j = 1, 2, \dots, m$) denote the assumption that the evolution and observation of the process at time t is defined by $X_t^{(j)}, V_t^{(j)}, G_t^{(j)}, X_t^{(j)}$. The Markovian transition of the process is described by $P(\mathcal{M}_t^{(j)}|\mathcal{M}_{t-1}^{(i)})$, the probability that the process swings to $\mathcal{M}_t^{(j)}$ from $\mathcal{M}_{t-1}^{(i)}$. A classic example of a system with a linear trend is to consider four models: the first is a default model which describes the system most of the time and three alternative models which represent an outlier observation, a step change and a slope change, by means of a suitable setting of $V_t^{(i)}$ and $W_t^{(i)}$, and which may explain a sudden general change in the behaviour of the time series.

We are interested in obtaining a formula analogous to (3.13). Following Harrison and Stevens (1971, 1976) we obtain,

$$P_t^{(j)} = P(\mathcal{M}_t^{(j)}|Y_t) = \sum_i P_t^{(i,j)}, \quad (3.14a)$$

where

$$P_t^{(i,j)} = P(M_t^{(j)}, M_{t-1}^{(i)} | Y_t) \propto p(Y_t | M_t^{(j)}, M_{t-1}^{(i)}) P(M_t^{(j)}, M_{t-1}^{(i)}), \quad \text{and} \quad (3.14b)$$

$$P(M_t^{(j)}, M_{t-1}^{(i)}) = P(M_t^{(j)} | M_{t-1}^{(i)}) P_{t-1}^{(i)}. \quad (3.14c)$$

Here $p(Y_t | M_t^{(j)}, M_{t-1}^{(i)})$ is given, again, either by (3.7b) or (3.10b) depending on whether Σ is known or not.

Although the cycle, in principle, can be carried on indefinitely, in practice the system rapidly becomes intractable even for a modest number of models. We can see the difficulty by noticing that for each $M_t^{(j)}$, the density $p(Y_t | M_t^{(j)}, M_{t-1}^{(i)})$ in (3.14b) depends on $M_{t-1}^{(i)}$ only through the prior $p(\Theta_t | M_t^{(j)})$ which is weighted by $P(M_{t-1}^{(i)})$. Hence, the posterior $p(\Theta_t | Y_t, M_t^{(j)})$ is weighted by $P(M_{t-1}^{(i)} | Y_t, M_t^{(j)}) = \frac{P_t^{(i,j)}}{P_t^{(j)}}$ and the total number of the posterior components grows geometrically.

Therefore, for the sake of tractability, a mixture-collapsing procedure similar to that for the DLM is required. Our criteria, based on a suggestion of Smith and West (1982), for approximating a mixture of densities by a single density is to minimize the Kullback and Liebler (1951) directed divergence.

In general, the problem of approximating the density of a random variable Z by a parametric density $p(Z|\Phi)$ using the Kullback-Liebler directed divergence as a criteria is equivalent to finding the optimal Φ in the following maximization problem:

$$\max_{\Phi} \mathbb{E}_Z \log p(Z|\Phi). \quad (3.15)$$

Here the expectation is taken over the target Z ; see Appendix A5.1.

As usual we consider first the case in which Σ is known. Our goal is to collapse, for each j , the posterior distribution $P(\Theta | Y_t, M_t^{(j)})$ which is a mixture of $N(M_t^{(i,j)}, C_t^{(i,j)}, \Sigma)$ with weights $\frac{P_t^{(i,j)}}{P_t^{(j)}}$ into a single $N(M_t^{(j)}, C_t^{(j)}, \Sigma)$. The distribution $P(\Theta_t | Y_t, M_t^{(j)}, M_{t-1}^{(i)})$ is $N(M_t^{(i,j)}, C_t^{(i,j)}, \Sigma)$ since we are assuming, of course, that the distribution $P(\Theta_{t-1} | M_{t-1}^{(i)})$ is $N(M_{t-1}^{(i)}, C_{t-1}^{(i)}, \Sigma)$.

In order to simplify the notation we solve first a more general problem, namely the approximation of the distribution of a random matrix Θ by a $N(M, C, \Sigma)$, where the free parameters are M and C . In other words, we want to obtain the optimal values for M and C in the maximization problem,

$$\max_{M, C} \mathbb{E}_{\Theta} L(\Theta | M, C) \quad (3.16)$$

where

$$L(\Theta | M, C) = -\frac{1}{2}pq \log(2\pi) - \frac{1}{2}q \log(|C|) - \frac{1}{2}p \log(|\Sigma|) - \frac{1}{2} \text{tr}((\Theta - M)' C^{-1} (\Theta - M) \Sigma^{-1}).$$

Differentiating $\mathbb{E}_{\Theta} L(\Theta | M, C)$ we obtain (see Press, 1982, p. 42-43, for the appropriate differentiating formulas),

$$\frac{\partial}{\partial M} \mathbb{E}_{\Theta} L(\Theta | M, C) = -\mathbb{E}_{\Theta} C^{-1} (\Theta - M) \Sigma^{-1} = -C^{-1} (\mathbb{E}_{\Theta} \Theta - M) \Sigma^{-1}, \quad (3.17a)$$

$$\frac{\partial}{\partial C^{-1}} \underset{\Theta}{\mathbb{E}} L(\Theta | M, C) = \underset{\Theta}{\mathbb{E}} \left(\frac{1}{2} q C - \frac{1}{2} (\Theta - M) \Sigma^{-1} (\Theta - M)' \right) = \frac{1}{2} (q C - \underset{\Theta}{\mathbb{E}} (\Theta - M) \Sigma^{-1} (\Theta - M)'). \quad (3.17b)$$

Therefore the optimal values M and C in (3.16) are given by,

$$\hat{M} = \underset{\Theta}{\mathbb{E}} \Theta \quad \text{and} \quad (3.18a)$$

$$\hat{C} = \frac{1}{q} \underset{\Theta}{\mathbb{E}} (\Theta - \hat{M}) \Sigma^{-1} (\Theta - \hat{M})' = \frac{1}{q} (\underset{\Theta}{\mathbb{E}} \Theta \Sigma^{-1} \Theta' - \hat{M} \Sigma^{-1} \hat{M}'). \quad (3.18b)$$

Note that \hat{C} is a symmetric positive definite matrix as required.

Applying (3.18) to our particular collapsing problem the following solution is obtained,

$$M_t^{(j)} = \sum_i \frac{P_t^{(i,j)}}{P_t^{(j)}} M_t^{(i,j)}, \quad (3.19a)$$

$$C_t^{(j)} = \frac{1}{q} \left(\sum_i \frac{P_t^{(i,j)}}{P_t^{(j)}} (q C_t^{(i,j)} + M_t^{(i,j)} \Sigma^{-1} M_t^{(i,j)'} - M_t^{(j)} \Sigma^{-1} M_t^{(j)'}) \right). \quad (3.19b)$$

The derivation of formulas (3.19) rests on two results. Firstly, the right-hand side expectations in (3.18) may be calculated using the fact that the expectation of a mixture is the mixtures of the expectations, e.g. $M_t^{(j)} = \underset{\Theta_t | Y_t, M_t^{(j)}}{\mathbb{E}} \Theta_t = \sum_{(i)} \left(\frac{P_t^{(i,j)}}{P_t^{(j)}} \right) \underset{\Theta_t | Y_t, M_t^{(j)}, M_{t-1}^{(i)}}{\mathbb{E}} \Theta_t$. Secondly, the resulting component expectations may be obtained using again (3.18) since the logarithmic scoring rule is proper (see Appendix A5.1), e.g. $M_t^{(i,j)} = \underset{\Theta_t | Y_t, M_t^{(j)}, M_{t-1}^{(i)}}{\mathbb{E}} \Theta_t$.

The equations (3.19), for $\Sigma = 1$, coincide with the DLM collapsing procedure of Harrison and Stevens (1971, 1976) because (3.19b) can be rewritten as

$$C_t^{(j)} = \sum_i \frac{P_t^{(i,j)}}{P_t^{(j)}} \left(C_t^{(i,j)} + \frac{1}{q} (M_t^{(i,j)} - M_t^{(j)}) \Sigma^{-1} (M_t^{(i,j)} - M_t^{(j)})' \right).$$

However, it does not coincide with the Smith and West (1983) procedure because Σ (c^2 in their notation) is missing in their formula for $C_t^{(j)}$. The implementation of (3.19) presents no difficulty provided that the number of the rival models is reasonably small. Notice that Σ^{-1} needs to be computed only once and the total computing time is at least proportional to the square of the number of possible models.

Let us now consider the case for Σ unknown. Our goal is to approximate a mixture of

$$\text{NW}^{-1}(M_t^{(i,j)}, C_t^{(i,j)}, S_t^{(i,j)}, d_t)$$

by a single $\text{NW}^{-1}(M_t^{(j)}, C_t^{(j)}, S_t^{(j)}, d_t)$. We are assuming as before that the prior, for each i , is distributed as a single $\text{NW}^{-1}(M_{t-1}^{(i)}, C_{t-1}^{(i)}, S_{t-1}^{(i)}, d_{t-1})$.

We start again with the more general problem of approximating the distribution of a random matrix $\begin{bmatrix} \Theta \\ \Sigma \end{bmatrix}$ where Σ is symmetric and positive definite, by a $\text{NW}^{-1}(M, C, S, d)$, where the free parameters are

M, C and S . In other words, we want to obtain the optimal values for M, C and S in the maximization problem,

$$\max_{M, C, S} \left[\begin{array}{c} \mathbb{E} \\ \Theta \\ \Sigma \end{array} \right] \log p(\Theta | M, C, S, d), \quad (3.20)$$

where $p(\Theta | M, C, S, d)$ is the density associated with the distribution $\text{NW}^{-1}(M, C, S, d)$. Using the matrix-normal inverted-Wishart decomposition of the density we can restate the maximization problem as,

$$\max_{M, C} \mathbb{E}_{\Sigma} \log p(\Theta | M, C, \Sigma) + \max_S \mathbb{E}_{\Sigma} \log p(\Sigma | S, d). \quad (3.21)$$

The problem of approximating the distribution of Σ by an $\text{W}^{-1}(S, d)$ is implicit in the second term of (3.21). Let $L(\Sigma | S, d)$ be the logarithm of the inverted-Wishart density (A3.2.15), then

$$\frac{\partial}{\partial S} \mathbb{E}_{\Sigma} L(\Sigma | S, d) = \frac{1}{2} \left((d + q - 1)S^{-1} - \mathbb{E}_{\Sigma} \Sigma^{-1} \right), \quad (3.22)$$

and the optimal value for S in (3.20) is given by,

$$\hat{S} = \nu \left(\mathbb{E}_{\Sigma} \Sigma^{-1} \right)^{-1}, \quad (3.23a)$$

where $\nu = d + q - 1$. From (3.17) clearly the optimal M and C for the first term in (3.21) must satisfy

$$O = \mathbb{E}_{\Sigma} \left(\hat{C}^{-1} \left(\mathbb{E}_{\Theta | \Sigma} \Theta - \hat{M} \right) \Sigma^{-1} \right)$$

and

$$O = \mathbb{E}_{\Sigma} \left(q \hat{C} - \mathbb{E}_{\Theta | \Sigma} (\Theta - \hat{M}) \Sigma^{-1} (\Theta - \hat{M})' \right),$$

i.e.

$$\hat{M} = \left(\begin{bmatrix} \mathbb{E} \\ \Theta \\ \Sigma \end{bmatrix} \Theta \Sigma^{-1} \right) \left(\mathbb{E}_{\Sigma} \Sigma^{-1} \right)^{-1} \quad \text{and} \quad (3.23b)$$

$$\hat{C} = \left[\begin{array}{c} \mathbb{E} \\ \Theta \\ \Sigma \end{array} \right] \frac{1}{q} (\Theta - \hat{M}) \Sigma^{-1} (\Theta - \hat{M}) = \frac{1}{q} \left(\begin{bmatrix} \mathbb{E} \\ \Theta \\ \Sigma \end{bmatrix} \Theta \Sigma^{-1} \Theta' - \hat{M} \left(\mathbb{E}_{\Sigma} \Sigma^{-1} \right) \hat{M}' \right). \quad (3.23c)$$

Therefore the optimal M, C and S for the problem (3.20) are given by (3.23). In our derivation of the optimal setting of M, C and S , we have not considered the constraints on C and S to be symmetric and positive definite, but the solution of the unconstrained optimization problem satisfies the constraints; observe (3.23a) and (3.23c). Thus, it is also the solution to the constrained problem.

Applying this result to our particular mixture-collapsing problem and using again the fact that the logarithmic scoring rule is proper, we readily obtain the solution,

$$S_t^{(j)} = \left(\sum_i \frac{P_t^{(i,j)}}{P_t^{(j)}} S_t^{(i,j)-1} \right)^{-1}, \quad (3.24a)$$

$$M_t^{(j)} = \left(\sum_i \frac{P_t^{(i,j)}}{P_t^{(j)}} M_t^{(i,j)} S_t^{(i,j)-1} \right) S_t^{(j)}, \quad (3.24b)$$

$$C_t^{(j)} = \frac{1}{q} \left(\left(\sum_i \frac{P_t^{(i,j)}}{P_t^{(j)}} (q C_t^{(i,j)} + M_t^{(j)} \nu S_t^{(j)-1} M_t^{(j)\prime}) \right) - M_t^{(j)} \nu S_t^{(j)-1} M_t^{(j)\prime} \right). \quad (3.24c)$$

For the DLM case, Σ being a scalar, (3.24a) coincides with the formula given by Smith and West (1983), but (3.24b) and (3.24c) do not. The reason is found in the missing factor pointed out in the comment after formulas (3.19).

In principle, a better approximation may be achieved by including d_t in the optimization procedure. Unfortunately, the resulting equations involve digamma functions, so that a time consuming numerical method is required for finding the solution. Moreover, even the implementation of (3.24) may be difficult when the dimension of Σ is large due to equation (3.24a).

It is important to notice that we have in equation (3.23a) implicitly assumed that $(d+q-1)$ is positive. Furthermore, throughout this section we have considered only non-singular distributions because for singular distributions the density function in (3.15) does not exist, or to put it in another way it involves Dirac delta functions. Therefore, it seems that there is no other way for dealing with singular distributions than to remove the redundancy; see (A3.2.20) and (3.29).

3.5 REPARAMETERIZATION.

Reparameterization concerns alternative but equivalent parametric representations of a dynamic model. The operation of a system can be more effective by employing the model representation most easily interpretable by the practitioner. In addition, a reparameterization may avoid numerical problems due to rounding errors.

An equivalent representation of (3.5), but in terms of $\Psi_t = H\Theta_t K$ and $\Xi = K'\Sigma K$ where H and K are non-singular is, according to (A3.2.5) and (A3.2.18), as follows.

Observation Equation:

$$(HY_t K) = (HX_t H^{-1})\Psi_t + (HE_t K), \quad (HE_t K) \sim N(O, HV_t H', \Xi), \quad (3.25a)$$

Evolution Equation:

$$\Psi_t = (HG_t H^{-1})\Psi_{t-1} + (HF_t K), \quad (HF_t K) \sim N(O, HW_t H', \Xi), \quad (3.25b)$$

Prior Information:

$$\begin{bmatrix} \Psi \\ \Xi \end{bmatrix} \sim NW^{-1}(HM_{t-1}K, HC_{t-1}H', K'S_{t-1}K, d_{t-1}). \quad (3.25c)$$

Inferences about the original parameters Θ and Σ can be made by means of the inverse expressions corresponding to (3.25c). For most applications, rescaling independent and dependent variables by taking $H = \text{diag}(h_1, \dots, h_p)$ and $K = \text{diag}(k_1, \dots, k_q)$ avoids numerical problems in the analysis. In

particular the setting $H = \lambda^{\frac{1}{2}}I$ and $K = \lambda^{-\frac{1}{2}}I$ leave the observation and evolution equations unchanged, $\Psi_t = \Theta_t$ and $\Xi = \lambda^{-1}\Sigma$ meaning that the effect produced by global rescaling of the driving variances V_t, W_t and C_{t-1} is completely absorbed by an inverse rescaling of the system scale variance.

3.6 SYSTEM DECOUPLING.

The DLMR model can be decoupled into several independent DLMR sub-models when the system scale variance is known and is block-diagonal. This result can be shown by looking at the system in its entirety. The DLMR model (3.6a), (3.6b) and (3.7b) can be rewritten as,

$$\begin{bmatrix} Y_t \\ \Theta_t \\ \Theta_{t-1} \end{bmatrix} = H_t \begin{bmatrix} E_t \\ F_t \\ \Theta_{t-1} \end{bmatrix}, \quad \begin{bmatrix} E_t \\ F_t \\ \Theta_{t-1} \end{bmatrix} \sim N \left(\begin{bmatrix} O \\ O \\ M_{t-1} \end{bmatrix}, \begin{bmatrix} V_t & O & O \\ O & W_t & O \\ O & O & C_{t-1} \end{bmatrix}, \Sigma \right), \quad (3.26)$$

where

$$H_t = \begin{bmatrix} I & X_t & X_t G_t \\ O & I & G_t \\ O & O & I \end{bmatrix}.$$

Thus, the joint distribution of $Y_t, \Theta_t, \Theta_{t-1}$ is,

$$\begin{bmatrix} Y_t \\ \Theta_t \\ \Theta_{t-1} \end{bmatrix} \sim N \left(H_t \begin{bmatrix} O \\ O \\ M_{t-1} \end{bmatrix}, H_t \begin{bmatrix} V_t & O & O \\ O & W_t & O \\ O & O & C_{t-1} \end{bmatrix} H_t', \Sigma \right). \quad (3.27)$$

Smoothing, filtering and prediction can be achieved by means of particular conditional and marginal distributions of (3.27). Incidentally, this representation suggests a straight-forward method for dealing with models with time correlated noises and/or correlated error and noise. This involves considering a non block-diagonal matrix for the left scale parameter in (3.26) and applying the appropriate formula from Appendix A3.2 (a different approach is taken in Chapter 7).

When the system scale variance is block-diagonal, say $\Sigma = \begin{bmatrix} \Sigma_{11} & O \\ O & \Sigma_{22} \end{bmatrix}$, the system (3.27) can be decoupled as several independent systems (the general result follows by induction):

$$\begin{bmatrix} Y_{jt} \\ \Theta_{jt} \\ \Theta_{j(t-1)} \end{bmatrix} \sim N \left(H_t \begin{bmatrix} O \\ O \\ M_{j(t-1)} \end{bmatrix}, H_t \begin{bmatrix} V_t & O & O \\ O & W_t & O \\ O & O & C_{t-1} \end{bmatrix} H_t', \Sigma_{jj} \right), \quad \text{for } j = 1, 2, \quad (3.28)$$

where $\Theta_t = [\Theta_{.1t}, \Theta_{.2t}]$ (preserving the conformability of $\Theta_t \Sigma \Theta_t'$), and Θ_{t-1}, M_t , and Y_t are partitioned in a similar way.

Let us assume now that Σ is unknown and is distributed in the usual manner, then the system joint distribution becomes,

$$\begin{bmatrix} Y_t \\ \Theta_t \\ \Theta_{t-1} \\ \Sigma \end{bmatrix} \sim NW^{-1} \left(H_t \begin{bmatrix} O \\ O \\ M_{t-1} \end{bmatrix}, H_t \begin{bmatrix} V_t & O & O \\ O & W_t & O \\ O & O & C_{t-1} \end{bmatrix} H_t', S_{t-1}, d_{t-1} \right). \quad (3.29)$$

Again, this representation suggests a way of handling models with non-zero covariances between E_t, F_t and Θ_{t-1} .

Given the additional information $H_0 : \Sigma_{12} = O$, in accordance with formula (A3.2.16), the system (3.29) can be decoupled into two independent models:

$$\begin{bmatrix} Y_{jt} \\ \Theta_{jt} \\ \Theta_{j(t-1)} \\ \Sigma_{jj} \end{bmatrix} \sim NW^{-1} \left(H_t \begin{bmatrix} O & O \\ O & O \\ M_{j(t-1)} \end{bmatrix}, H_t \begin{bmatrix} V_t & O & O \\ O & W_t & O \\ O & O & C_{t-1} \end{bmatrix} H_t', S_{jj(t-1)}, d_{j(t-1)} \right), \quad \text{for } j = 1, 2, \quad (3.30)$$

where $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, S_{t-1} is partitioned in a similar way, $d_{j(t-1)} = d_{t-1} + 2(q - q_j)$ and q_j is the dimension of Σ_{jj} .

Therefore, we can use the multi-process class I approach and apply Bayes' theorem in order to test the null hypothesis H_0 against the alternative, since the predictive density of Y_t under the null model (the original model given H_0), is simply the product of the predictive densities of Y_{1t} and Y_{2t} , and is obtainable from the independent modes (3.30). This is, in fact, the Jeffreys (1961) general method for testing a null hypothesis H_0 . An example is given at the end of Chapter 5. The extension of the method for dealing with the hypothesis that Σ is block-diagonal is easily appreciated from (3.30).

3.7 DYNAMIC WEIGHTED MULTIVARIATE REGRESSION.

When the observations of a DLMR are row vectors, $r = 1$ in (3.5), the model is referred to as a DWMR. As mentioned before, this is a dynamic formulation of the familiar multivariate regression model, and it is very useful for modelling multiple time series that present similar dynamics. Two examples involving real data are discussed in the following chapters, here we restrict ourselves to some theoretical aspects.

According to (3.5) the DWMR is defined by,

Observation Equation:

$$\underline{y}'_t = \underline{x}'_t \Theta_t + \underline{\epsilon}'_t, \quad \underline{\epsilon}'_t \sim N(\underline{0}, v_t \Sigma). \quad (3.31a)$$

Evolution Equation:

$$\Theta_t = G_t \Theta_{t-1} + F_t, \quad F_t \sim N(O, W_t, \Sigma). \quad (3.31b)$$

Prior Information:

$$\Theta_{t-1} \sim N(M_{t-1}, C_{t-1}, \Sigma), \quad \Sigma \sim W^{-1}(S_{t-1}, d_{t-1}). \quad (3.31c)$$

Where $\underline{\epsilon}_t$, F_t and Θ_{t-1} are independent given Σ . Its updating formula for the evolution remains as in (3.10a), but the prediction and posterior equations are reduced from (3.10) to,

Prediction:

$$\underline{y}_t \sim t(\hat{y}_t, \check{y}_t S_{t-1}, d_{t-1}), \quad \text{where } \check{y}_t = v_t + \underline{x}'_t C_t^* \underline{x}_t \quad \text{and } \hat{y}_t = M'_{t-1} G'_t \underline{x}_t. \quad (3.32a)$$

Posterior:

$$\begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix} | \underline{y}_t \sim \text{NW}^{-1}(M_t, C_t, S_t, d_t), \quad \text{where } M_t = G_t M_{t-1} + \underline{a}_t \hat{\epsilon}'_t, \quad C_t = C_t^* - \check{y}_t \underline{a}'_t \underline{a}_t, \\ S_t = S_{t-1} + \frac{1}{\check{y}_t} \hat{\epsilon}_t \hat{\epsilon}'_t, \quad d_t = d_{t-1} + 1, \quad \hat{\epsilon}_t = \underline{y}_t - \hat{y}_t \quad \text{and } \underline{a}_t = \frac{1}{\check{y}_t} C_t^* \underline{x}_t. \quad (3.32b)$$

The notation for the multivariate \underline{t} distribution is discussed at the end of Appendix A3.2. Note that these updating formulas require no matrix inversion and in consequence they are very easy to implement.

The vector representation (3.12) corresponding to (3.31) is,

Observation Equation:

$$\underline{y}_t = (I \otimes \underline{x}'_t) \underline{\theta}_t + \underline{\varepsilon}_t, \quad \underline{\varepsilon}_t \sim N(0, v_t \Sigma). \quad (3.33a)$$

Evolution Equation:

$$\underline{\theta}_t = (I \otimes G_t) \underline{\theta}_{t-1} + \underline{f}_t, \quad \underline{f}_t \sim N(0, \Sigma \otimes W_t). \quad (3.33b)$$

Prior Information:

$$\underline{\theta}_{t-1} \sim N(\underline{m}_{t-1}, \Sigma \otimes C_{t-1}), \quad \Sigma \sim W^{-1}(S_{t-1}, d_{t-1}). \quad (3.33c)$$

Where $\underline{\theta}_t = \text{vec } \Theta_t$, etc. Both representations (3.31) and (3.33) can be reformulated in terms of the individual models for each time series,

$$y_{jt} = \underline{x}'_t \underline{\theta}_{jt} + e_{jt}, \quad e_{jt} \sim N(0, \sigma_{jj} v_t), \quad j = 1, \dots, q \quad (3.34a)$$

$$\underline{\theta}_{jt} = G_t \underline{\theta}_{j(t-1)} + \underline{f}_{jt}, \quad \underline{f}_{jt} \sim N(0, \sigma_{jj} W_t), \quad j = 1, \dots, q \quad (3.34b)$$

$$\underline{\theta}_{j(t-1)} \sim N(\underline{m}_{j(t-1)}, \sigma_{jj} C_{(t-1)}), \quad j = 1, \dots, q \quad (3.34c)$$

where $\underline{y}'_t = [y_{1t}, \dots, y_{qt}]$, $\Theta_t = [\underline{\theta}_{1t}, \dots, \underline{\theta}_{qt}]$, etc. These models are linked by an additional assumption; e_{jt}, f_{jt} and $\Theta_{ij(t-1)}$ are mutually independent (given Σ), distributed as multivariate normals with $\text{cov}(e_{jt}, e_{j't}) = \sigma_{jj'} v_t$, $\text{cov}(f_{ijt}, f_{i'j't}) = \sigma_{jj'} W_{ii'j't}$, $\text{cov}(\Theta_{ij(t-1)}, \Theta_{i'j'(t-1)}) = C_{ii'j'(t-1)} \sigma_{jj'}$ and $\Sigma \sim W^{-1}(S_{t-1}, d_{t-1})$.

Equations (3.31a) and (3.31c) define the usual likelihood and prior distributions for the multivariate regression model with weights $\frac{1}{v_t}$ and common regressors. Equation (3.31b) introduces the dynamic structure to the model. From these equations it is clear that the observational error, and each row of the regression parameters Θ_t have essentially the same variance-covariance structure given by Σ , i.e. they have the same variance matrix except for a scalar factor. Furthermore, this structure is preserved in the evolution because the rows of the evolution noise F_t have essentially the same variance-covariance structure. It is also apparent from (3.34) that marginally each series y_{jt} is being modelled as a dynamic weighted univariate regression, where the regression coefficients $\underline{\theta}_{jt}$ given by the columns of Θ associated with each y_{jt} , are evolving in a similar way in the sense that they share the same evolution matrix G_t .

and have essentially the same noise variance-covariance structure given by W_t . It is precisely this special structure of the DWMR that keeps the analysis tractable when Σ is unknown.

Besides significant savings in computing time, the reason for considering a multivariate DWMR instead of a set of univariate DWMR's, is that it provides the on-line procedure necessary for learning about the full variance-covariance structure Σ in order to make proper joint forecasts for y_{1t}, \dots, y_{qt} . This permits us, among other things, to update the joint forecast of a subset of y_t 's when the values of another subset are observed. A simple procedure based on the sweep-operator for obtaining these contemporaneous conditional predictive distributions is described in Section 4.2.

Comparing the representation (3.33) of the DWMR and the formulation of the DLM or the extended DLM (see Section 3.3), it is evident that the DLM has a more general structure. On the other hand, the DWMR offers more freedom for learning about the variance-covariance structure of the system. Therefore, when we consider the use of a DWMR instead of a DLM for modelling a multivariate series, we are trading generality of the model for freedom in the learning procedure about the system variance-covariance structure. In many cases the price is worth paying, typical examples are given in chapters 5, 6 and 7.

The DWMR was developed in Quintana (1985) and extended to allow for dynamic scale variances in Quintana and West (1986). This latter model is discussed in Section 6.3. Independently, a non-Bayesian formulation, essentially (3.33) with a vague-like prior (the details can be seen in Section 4.1), was presented in Harvey (1986). And also independently, a Bayesian non-weighted ($v_t = 1$) steady ($G_t = I$) version of the DWMR was formulated by Highfield (1984).

APPENDIX A3.1.

KRONECKER DIRECT PRODUCT AND vec OPERATOR.

The Kronecker direct product and vec operator, in combination with the standard matrix theory, play a central role in the theory of the DLMR. The necessary definitions and results are included here for convenience. Additional information and sketch proofs may be found in Searle (1982) and Press (1982). Furthermore, it is assumed that the reader is familiar with basic results of the trace function such as: $\text{tr}(AB) = \text{tr}(BA)$, $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$, $\text{tr}(\alpha A) = \alpha \text{tr } A$, etc.; see for example Press (1982, p. 31-33).

A3.1.1 Kronecker Direct Product.

(a) Notation:

$A \otimes B$ denotes the Kronecker direct product between the matrices A and B , i.e.

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1m}B \\ \vdots & \ddots & \vdots \\ a_{n1}B & \dots & a_{nm}B \end{bmatrix}, \quad (\text{A3.1.1})$$

where A is a $(n \times m)$ matrix given by $A = [a_{ij}]$.

(b) Block diagonal:

$$I \otimes A = \text{diag}(A, \dots, A). \quad (\text{A3.1.2})$$

(c) Matrix product:

$$(A \otimes B)(C \otimes F) = AC \otimes BF \quad (\text{A3.1.3})$$

(it is implicitly assumed that the products are defined).

(d) Inverse:

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}. \quad (\text{A3.1.4})$$

(e) Determinant:

$$|A \otimes B| = |A|^n |B|^m, \quad (\text{A3.1.5})$$

where A is $(m \times m)$ and B is $(n \times n)$.

(f) Factorization:

$$A \otimes (B + C) = A \otimes B + A \otimes C \quad \text{and} \quad (A + B) \otimes C = A \otimes C + B \otimes C. \quad (\text{A3.1.6})$$

(g) Transpose:

$$(A \otimes B)' = A' \otimes B'. \quad (\text{A3.1.7})$$

(h) Trace:

$$\text{tr}(A \otimes B) = (\text{tr } A)(\text{tr } B), \quad (\text{A3.1.8})$$

where A and B are square matrices.

(i) Scale:

$$\alpha(A \otimes B) = (\alpha A) \otimes B = A \otimes (\alpha B), \quad (\text{A3.1.9a})$$

$$\alpha \otimes A = A \otimes \alpha, \quad (\text{A3.1.9b})$$

where α is a scalar.

(j) Partition:

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \otimes B = \begin{bmatrix} A_{11} \otimes B & A_{12} \otimes B \\ A_{21} \otimes B & A_{22} \otimes B \end{bmatrix}. \quad (\text{A3.1.10})$$

(k) Eigen Structure:

The spectral decomposition of $A \otimes B$ is,

$$A \otimes B = (P \otimes Q)(\Lambda \otimes \Gamma)(P \otimes Q)^{-1}, \quad (\text{A3.1.11})$$

where $A = P\Lambda P^{-1}$ and $B = Q\Gamma Q^{-1}$ are the spectral decompositions of A and B .

(l) Square Root:

Let A and B be non-negative definite symmetric matrices, then,

$$(A \otimes B)^{\frac{1}{2}} = A^{\frac{1}{2}} \otimes B^{\frac{1}{2}}, \quad (\text{A3.1.12})$$

where $\frac{1}{2}$ denotes the symmetric square root of a matrix i.e. if the spectral decomposition of A is $A = P\Lambda P'$ then $A^{\frac{1}{2}} = P\Lambda^{\frac{1}{2}}P'$, where $\Lambda^{\frac{1}{2}} = \text{diag}(\lambda_1^{\frac{1}{2}}, \dots, \lambda_n^{\frac{1}{2}})$.

A3.1.2 The vec Operator.

(a) Notation:

$\text{vec } A$ denotes the usual column-vectorization of the matrix A , i.e.

$$\text{vec } A = [\underline{a}'_1, \dots, \underline{a}'_m]', \quad (\text{A3.1.13})$$

where $A = [a_1, \dots, a_m]$.

(b) Sum:

$$\text{vec}(A + B) = \text{vec } A + \text{vec } B. \quad (\text{A3.1.14})$$

(c) Scale:

$$\text{vec}(\alpha A) = \alpha \text{vec } A, \quad (\text{A3.1.15})$$

where α is a scalar.

(d) Product:

$$\text{vec } AXB = (B' \otimes A) \text{vec } X. \quad (\text{A3.1.16})$$

(e) Trace:

$$\text{tr}(AB) = \text{tr}((\text{vec } B)(\text{vec } A')') = (\text{vec } A')'(\text{vec } B), \quad (\text{A3.1.17a})$$

$$\text{tr}(ABCD) = (\text{vec } A')'(D' \otimes B)(\text{vec } C), \quad (\text{A3.1.17b})$$

where AB and $ABCD$ are square matrices.

APPENDIX A3.2.

BASIC MATRIX-VARIATE DISTRIBUTION THEORY.

It is safe to say that the study of the DLMR is equivalent to that of the matrix-normal inverted-Wishart distribution. The main results of the general (singular and non-singular) matrix-normal inverted-Wishart distribution are developed below. Proofs of the properties employed in the derivation of the DLMR updating recurrences (3.7) and (3.10) are included for completeness. The choice of a suitable notation is crucial; ours is essentially the notation of Box and Tiao (1973, Chapter 8), later extended by Dawid (1981) to the singular case.

A3.2.1 Matrix-normal Distribution.

The definition and properties of the matrix-normal distribution can be derived from those of the multivariate normal distribution; however, a simple and direct approach is preferred here. By so doing, the definition and properties of the multivariate normal distribution are obtained simultaneously.

Let Ξ be a random matrix such that its components are distributed independently as standard normals. Clearly, its characteristic function is given by,

$$f_{\Xi}(T) = \mathbb{E}_{\Xi} \exp(i \operatorname{tr}(T' \Xi)) = \prod_{k,j} \exp(it_{kj} \xi_{kj}) = \exp(-\frac{1}{2} \operatorname{tr}(T'T)). \quad (\text{A3.2.1})$$

Moreover, the characteristic function of $\Theta = H\Xi K + M$ is,

$$f_{\Theta}(T) = \mathbb{E}_{\Theta} \exp(i \operatorname{tr}(T' \Theta)) = \exp(-\frac{1}{2} \operatorname{tr}(T' C T \Sigma) + i \operatorname{tr}(T' M)), \quad (\text{A3.2.2})$$

where $C = HH'$ and $\Sigma = K'K$. This result may be verified as follows:

$$\mathbb{E}_{\Theta} \exp(i \operatorname{tr}(T' \Theta)) = \exp(i \operatorname{tr}(T' M)) \mathbb{E}_{\Xi} \exp(i \operatorname{tr}(KT' H\Xi)),$$

but from (A3.2.1)

$$\mathbb{E}_{\Xi} \exp(i \operatorname{tr}(KT' H\Xi)) = \exp(-\frac{1}{2} \operatorname{tr}(KT' H(KT' H)')) = \exp(-\frac{1}{2} \operatorname{tr}(T' C T \Sigma)).$$

Furthermore, the characteristic function of Θ depends on H and K only through C and Σ and given C and Σ non-negative definite symmetric matrices there always are H and K such that $C = HH'$ and $\Sigma = K'K$. Therefore the definition given below is admissible.

(a) Notation:

$$\Theta \sim N(M, C, \Sigma) \quad \text{if and only if } \mathbb{E}_{\Theta} \exp(i \operatorname{tr}(T' \Theta)) = \exp(-\frac{1}{2} \operatorname{tr}(T' C T \Sigma) + i \operatorname{tr}(T' M)), \quad (\text{A3.2.3})$$

where M is arbitrary and C and Σ are non-negative definite symmetric matrices. We say that Θ is distributed as a matrix normal with mean M and scale (variance) parameters C and Σ . This

representation is not minimal (O'Hagan, 1972), i.e. $\text{N}(M, C, \Sigma)$ and $\text{N}(M, \lambda C, \frac{1}{\lambda} \Sigma)$ where $\lambda > 0$ denote the same distribution. Notice that, if either $C = O$ or $\Sigma = O$ then Θ is a degenerate random matrix which has its probability concentrated in M . For $\Sigma = I$, the vector $\underline{\theta}$ follows the multivariate normal distribution denoted by $\text{N}(\underline{m}, C)$.

(b) Transpose:

$$\Theta \sim \text{N}(M, C, \Sigma) \quad \text{if and only if } \Theta' \sim \text{N}(M', \Sigma, C). \quad (\text{A3.2.4})$$

This follows since

$$f_{\Theta'}(T) = \underset{\Theta}{\mathbb{E}} \exp(i \operatorname{tr}(T\Theta)) = \exp(-\frac{1}{2} \operatorname{tr}(TCT'\Sigma) + i \operatorname{tr}(TM)) = \exp(-\frac{1}{2} \operatorname{tr}(T'\Sigma TC) + i \operatorname{tr}(T'M'))$$

and $\Theta = \Theta''$.

(c) Linear Transformation:

$$\Theta \sim \text{N}(M, C, \Sigma) \quad \text{implies } H\Theta K + L \sim \text{N}(HMK + L, HCH', K'\Sigma K). \quad (\text{A3.2.5})$$

This follows since the characteristic function of $H\Theta K + L$ is given by,

$$\begin{aligned} f_{H\Theta K + L}(T) &= \exp(i \operatorname{tr}(T'L)) \underset{\Theta}{\mathbb{E}} \exp(i \operatorname{tr}(KT'H\Theta)) = \\ &= \exp(i \operatorname{tr}(T'(L + HMK))) \exp(-\frac{1}{2} T' HCH' TK' \Sigma K). \end{aligned}$$

(d) Standard Representation:

Given any M, C and Σ there is $\Xi \sim \text{N}(O, I, I)$ such that

$$C^{\frac{1}{2}} \Xi \Sigma^{\frac{1}{2}} + M \sim \text{N}(M, C, \Sigma), \quad (\text{A3.2.6})$$

where $A^{\frac{1}{2}}$ denotes the symmetric square root of A as in (A3.1.12).

(e) Equivalence with the multivariate normal:

$$\Theta \sim \text{N}(M, C, \Sigma) \quad \text{if and only if } \operatorname{vec} \Theta \sim \text{N}(\operatorname{vec} M, \Sigma \otimes C). \quad (\text{A3.2.7})$$

Using (A3.1.15) clearly the characteristic function of $\operatorname{vec} \Theta$ is,

$$\begin{aligned} f_{\operatorname{vec} \Theta}(\operatorname{vec} T) &= \underset{\Theta}{\mathbb{E}} \exp(i \operatorname{tr}(T'\Theta)) = \exp(i \operatorname{tr}(T'M)) \exp(-\frac{1}{2} \operatorname{tr}(T'CT\Sigma)) \\ &= \exp(i(\operatorname{vec} T)' \operatorname{vec} M) \exp(-\frac{1}{2}(\operatorname{vec} T)' (\Sigma \otimes C) \operatorname{vec} T). \end{aligned}$$

(f) First and second moments:

$$\Theta \sim \text{N}(M, C, \Sigma) \quad \text{implies } \underset{\Theta}{\mathbb{E}} \operatorname{vec} \Theta = \operatorname{vec} M \quad \text{and } \underset{\Theta}{\text{VAR}} \operatorname{vec} \Theta = \Sigma \otimes C. \quad (\text{A3.2.7a})$$

From (A3.2.6) we have,

$$\begin{aligned}\mathbb{E}_{\Theta} \text{vec } \Theta &= \mathbb{E}_{\Xi} \text{vec}(C^{\frac{1}{2}} \Xi \Sigma^{\frac{1}{2}} + M) = \mathbb{E}_{\Xi}((\Sigma^{\frac{1}{2}} \otimes C^{\frac{1}{2}}) \text{vec } \Xi + \text{vec } M) = \text{vec } M, \\ \text{VAR}_{\Theta} \text{vec } \Theta &= \text{VAR}_{\Xi}((\Sigma^{\frac{1}{2}} \otimes C^{\frac{1}{2}}) \text{vec } \Xi + \text{vec } M) = \Sigma \otimes C.\end{aligned}$$

Moreover it is not difficult to see that (A3.2.7a) can be restated as

$$\begin{aligned}\Theta \sim N(M, C, \Sigma) \quad \text{implies} \\ \mathbb{E}_{\Theta} \Theta = M, \quad \mathbb{E}_{\Theta}(\Theta - M)(\Theta - M)' = C \text{tr } \Sigma \quad \text{and} \quad \mathbb{E}_{\Theta}(\Theta - M)'(\Theta - M) = \Sigma \text{tr } C.\end{aligned}\tag{A3.2.7b}$$

(g) Marginal and conditional distributions:

$$\begin{aligned}\Theta \sim N(M, C, \Sigma) \quad \text{if and only if} \\ \Theta_{1.} \sim N(M_{1.}, C_{11}, \Sigma) \quad \text{and} \quad \Theta_{2.} | \Theta_{1.} \sim N(M_{2.1.}, C_{22|1.}, \Sigma),\end{aligned}\tag{A3.2.8}$$

where $\Theta = \begin{bmatrix} \Theta_{1.} \\ \Theta_{2.} \end{bmatrix}$, $C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$ (preserving the conformability of $\Theta' C \Theta$), $M_{2.1.} = M_{2.} + C_{21} C_{11}^{-1} E_{1.}$, $E_{1.} = \Theta_{1.} - M_{1.}$, $\begin{bmatrix} M_{1.} \\ M_{2.} \end{bmatrix} = M$ and $C_{22|1.} = C_{22} - C_{21} C_{11}^{-1} C_{12}$ (C_{11}^{-1} is assumed to exist). Furthermore $\Theta_{1.}$ and $E_{2.1.} = \Theta_{2.} - M_{2.1.}$ are independent. A similar result for subsets of columns may be obtained readily in view of (A3.2.4).

We can verify (A3.2.8) directly by construction. Let $E_{2.1.} = \Theta_{2.} - M_{2.1.}$, then using (A3.2.5), it is not difficult to see that,

$$\begin{bmatrix} \Theta_{1.} \\ E_{2.1.} \end{bmatrix} = \begin{bmatrix} I & O \\ -C_{21} C_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} \Theta_{1.} \\ \Theta_{2.} \end{bmatrix} + \begin{bmatrix} O \\ C_{21} C_{11}^{-1} M_{1.} - M_{2.} \end{bmatrix} \sim N\left(\begin{bmatrix} M_{1.} \\ O \end{bmatrix}, \begin{bmatrix} C_{11} & O \\ O & C_{22|1.} \end{bmatrix}, \Sigma\right).$$

But (A3.2.3) implies that the characteristic function of $\begin{bmatrix} \Theta_{1.} \\ E_{2.1.} \end{bmatrix}$ is

$$\begin{aligned}f_{\begin{bmatrix} \Theta_{1.} \\ E_{2.1.} \end{bmatrix}}\left(\begin{bmatrix} T_{1.} \\ T_{2.} \end{bmatrix}\right) &= \exp(i \text{tr}(T_1' M_{1.})) \exp(-\frac{1}{2} \text{tr}(T_1' C_{11} T_{1.}, \Sigma)) \exp(-\frac{1}{2} \text{tr}(T_2' C_{22|1.} T_{2.}, \Sigma)) \\ &= f_{\Theta_{1.}}(T_{1.}) f_{E_{2.1.}}(T_{2.}),\end{aligned}$$

i.e. $\Theta_{1.}$ and $E_{2.1.}$ are distributed independently, $\Theta_{1.} \sim N(M_{1.}, C_{11}, \Sigma)$, $E_{2.1.} \sim N(O, C_{22|1.}, \Sigma)$ and $\Theta_{2.} = E_{2.1.} + M_{2.1.} \sim N(M_{2.1.}, C_{22|1.}, \Sigma)$ given $\Theta_{1.}$.

It is important to notice that marginally $\Theta_{1.} \sim N(M_{1.}, C_{11}, \Sigma)$ even if C_{11} is singular. Moreover, the conditional distribution of $\Theta_{2.} | \Theta_{1.}$ always can be obtained by applying (A3.2.8) sequentially row-by-row and skipping the conditionally degenerated rows, because such degenerated rows provide no further information. Finally, $\Theta_{1.}$ and $\Theta_{2.}$ are independent if and only if $C_{12} = O$.

(h) Non-singular matrix normal density:

A matrix-normal random variable is non-singular if and only if both scale parameter matrices are positive definite. In this case,

$$\Theta \sim N(M, C, \Sigma) \quad \text{if and only if} \quad C^{-\frac{1}{2}}(\Theta - M)\Sigma^{-\frac{1}{2}} \sim N(O, I, I).\tag{A3.2.9}$$

The density of Θ is given by,

$$p(\Theta) = k(C, \Sigma) \exp(-\frac{1}{2} \text{tr}(\Theta - M)' C^{-1} (\Theta - M) \Sigma^{-1}), \quad (\text{A3.2.10})$$

where $k(C, \Sigma) = (2\pi)^{-\frac{1}{2}pq} |C|^{-\frac{1}{2}q} |\Sigma|^{-\frac{1}{2}p}$, p and q are the number of rows and columns of Θ .

The result (A3.2.9) is an immediate consequence of (A3.2.5), and (A3.2.10) follows after a little algebra using results from Appendix A3.1 together with (A3.2.9), (A3.2.7) and the standard transformation formula.

A3.2.2 Inverted-Wishart Distribution.

Following Dawid (1981), the inverted-Wishart distribution can be defined as follows.

(a) Notation:

$$\Sigma \sim W^{-1}(S, d) \quad \text{if and only if } \Sigma \text{ has the same distribution as that of } A' \Omega_{(n,d)} A, \quad (\text{A3.2.11})$$

where $S = A'A$ and $\Omega_{(n,d)}$ is a $n \times n$ random positive definite symmetric matrix with probability density proportional to $|\Omega|^{-(\frac{1}{2}d+n)} \exp(-\frac{1}{2} \text{tr}(\Omega^{-1}))$ and $d > 0$. The matrix A and n need not be further specified because if a matrix B is such that $B'B = A'A$ then $A' \Omega_{(n,d)} A$ and $B' \Omega_{(N,d)} B$ have the same distribution; see reference above for details.

(b) Linear Transformation:

$$\Sigma \sim W^{-1}(S, d) \quad \text{implies } K' \Sigma K \sim W^{-1}(K' S K, d). \quad (\text{A3.2.12})$$

The matrix K is not necessarily square. From definition (A3.2.11)

$$K' \Sigma K \sim K' A' \Omega_{(n,d)} A K \sim W^{-1}(K' S K, d).$$

(c) Marginal Distribution:

$$\Sigma \sim W^{-1}(S, d) \quad \text{implies } \Sigma_{11} \sim W^{-1}(S_{11}, d), \quad (\text{A3.2.13})$$

where $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ and $S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$. This result is a direct consequence of (A3.2.12).

Mean:

$$\mathbb{E} \Sigma = \frac{S}{d-2}, \quad d > 0. \quad (\text{A3.2.14})$$

From (A3.2.11) we have, $\mathbb{E} \Sigma = A' \mathbb{E}_{\Omega_{(n,d)}} \Omega_{(n,d)} A = \frac{S}{d-2}$ since $\mathbb{E}_{\Omega_{(n,d)}} \Omega_{(n,d)} = \frac{I}{d-2}$; see for example Press (1982, p. 119).

(d) Non-singular inverted-Wishart:

An inverted-Wishart random variable $\Sigma_{(q \times q)}$ is non-singular if and only if the scale parameter S is non-singular. In this case its probability density is given by,

$$p(\Sigma) = k(S, d) |\Sigma|^{-(\frac{1}{2}d+q)} \exp(-\frac{1}{2} \text{tr}(S\Sigma^{-1})), \quad (\text{A3.2.15})$$

where

$$(k(S, d))^{-1} = 2^{\frac{1}{2}q\nu} \pi^{\frac{1}{4}q(q-1)} \prod_{j=1}^q (\frac{1}{2}(\nu - j - 1))! |S|^{-\frac{1}{2}\nu},$$

$\nu = d + q - 1$ and ! denotes the extended factorial; see Section 1.5.

From definition (A3.2.11) and the standard transformation formula it is clear that

$$\begin{aligned} p(\Sigma) &\propto |A'^{-1}\Sigma A^{-1}|^{-(\frac{1}{2}d+q)} \exp(-\frac{1}{2} \text{tr}((A'^{-1}\Sigma A^{-1})^{-1})) \text{abs}\left(\left|\frac{\partial A'^{-1}\Sigma A^{-1}}{\partial \Sigma}\right|\right) \\ &\propto |\Sigma|^{-(\frac{1}{2}d+q)} \exp(-\frac{1}{2} \text{tr}(S\Sigma^{-1})), \end{aligned}$$

i.e. Σ follows, in fact, the usual (non-singular) inverted-Wishart distribution, e.g. Box and Tiao (1973, p. 460).

(e) Conditional distribution given $\Sigma_{12} = O$:

$$\begin{aligned} \Sigma \sim W^{-1}(S, d) \text{ (non-singular) implies that given } \Sigma_{12} = O, \\ \Sigma_{11} \sim W^{-1}(S_{11}, d + 2q_2) \quad \text{and} \quad \Sigma_{22} \sim W^{-1}(S_{22}, d + 2q_1) \quad \text{independently,} \end{aligned} \quad (\text{A3.2.16})$$

where Σ and S are partitioned as in (A3.2.13), and q_1, q_2 are the dimensions of Σ_{11} and Σ_{22} .

This result follows since,

$$\begin{aligned} p(\Sigma_{11}, \Sigma_{22} | \Sigma_{12} = O) &\propto \begin{vmatrix} \Sigma_{11} & O \\ O & \Sigma_{22} \end{vmatrix}^{-(\frac{1}{2}d+q)} \exp\left(-\frac{1}{2} \text{tr}\left(S \begin{bmatrix} \Sigma_{11}^{-1} & O \\ O & \Sigma_{22}^{-1} \end{bmatrix}\right)\right) \\ &= |\Sigma_{11}|^{-(\frac{1}{2}(d+2q_2)+q_1)} \exp(-\frac{1}{2} \text{tr}(S_{11}\Sigma_{11}^{-1})) |\Sigma_{22}|^{-(\frac{1}{2}(d+2q_1)+q_2)} \exp(-\frac{1}{2} \text{tr}(S_{22}\Sigma_{22}^{-1})). \end{aligned}$$

A3.2.3 Matrix-normal Inverted-Wishart Distribution.

The matrix-normal inverted-Wishart distribution may be defined as follows.

(a) Notation:

$$\begin{bmatrix} \Theta \\ \Sigma \end{bmatrix} \sim NW^{-1}(M, C, S, d) \quad \text{if and only if } \Theta \sim N(M, C, \Sigma) \text{ given } \Sigma \quad \text{and} \quad \Sigma \sim W^{-1}(S, d). \quad (\text{A3.2.17})$$

(b) Linear transformation:

$$\begin{bmatrix} \Theta \\ \Sigma \end{bmatrix} \sim NW^{-1}(M, C, S, d) \quad \text{implies} \quad \begin{bmatrix} H\Theta K + L \\ K'\Sigma K \end{bmatrix} \sim NW^{-1}(HMK + L, HCH', K'SK, d). \quad (\text{A3.2.18})$$

This result follows from definition (A3.2.17) using (A3.2.5) and (A3.2.12).

(c) Marginal and conditional distributions:

The marginal distribution of Θ in definition (A3.2.17) is denoted as $\Theta \sim T(M, C, S, d)$ and referred to as the matrix-T distribution; see Section A3.2.4.

The conditional distribution of Σ given Θ follows,

$$\Sigma \sim W^{-1}(S + (\Theta - M)'C^{-1}(\Theta - M), d + p), \quad (A3.2.19)$$

where p is the number of rows of Θ .

First, we derive (A3.2.19) for S non-singular. Formulas (A3.2.10) and (A3.2.15) imply,

$$p(\Sigma|\Theta) \propto |\Sigma|^{-(\frac{1}{2}(d+p)+q)} \exp(-\frac{1}{2} \text{tr}((S + (\Theta - M)'C^{-1}(\Theta - M))\Sigma^{-1})),$$

and the result follows from (A3.2.15).

On the other hand, if S is positive semidefinite then we can assume, without loss of generality, that $S = [I, S_{11}^{-1}S_{12}]'S_{11}[I, S_{11}^{-1}S_{12}]$, where S_{11} is positive definite and the rank of S_{11} is the same as that of S ; see formula A4.1.11. Thus, according to (A3.2.18), $\begin{bmatrix} \Theta \\ \Sigma \end{bmatrix}$ can be written as,

$$\begin{bmatrix} \Theta \\ \Sigma \end{bmatrix} = \begin{bmatrix} M + (\Theta_{.1} - M_{.1})[I, S_{11}^{-1}S_{12}] \\ [I, S_{11}^{-1}S_{12}]'\Sigma_{11}[I, S_{11}^{-1}S_{12}] \end{bmatrix}, \quad (A3.2.20)$$

where $\Theta = [\Theta_{.1}, \Theta_{.2}]$, $M = [M_{.1}, M_{.2}]$ and $\begin{bmatrix} \Theta_{.1} \\ \Sigma_{11} \end{bmatrix} \sim NW^{-1}(M_{.1}, C, S_{11}, d)$.

But $\Sigma_{11} \sim W^{-1}(S_{11} + (\Theta_{.1} - M_{.1})'C^{-1}(\Theta_{.1} - M_{.1}), d + p)$ given $\Theta_{.1}$, and (A3.2.20) imply $\Sigma \sim W^{-1}(S + (\Theta - M)'C^{-1}(\Theta - M), d + p)$ given Θ as claimed in (A3.2.19).

Formula (A3.2.19) is, of course, meaningless when C is singular, but the conditional distribution of Σ given Θ always can be found by applying sequentially the formula below; see the relevant comment following (A3.2.8). Thus

$$\begin{aligned} \begin{bmatrix} \Theta \\ \Sigma \end{bmatrix} &\sim NW^{-1}(M, C, S, d) \quad \text{if and only if} \\ \Theta_{.1} &\sim T(M_{.1}, C_{11}, S, d) \quad \text{and} \quad \begin{bmatrix} \Theta_{.2} \\ \Sigma \end{bmatrix} \sim NW^{-1}(M_{2.|1.}, C_{22|1.}, S_{|1.}, d + p_{1.}) \quad \text{given } \Theta_{.1}, \end{aligned} \quad (A3.2.21)$$

where $\Theta, M, E_{1.}, M_{2.|1.}, C_{22|1.}$ are as in (A3.2.8), $p_{1.}$ is the number of rows of $\Theta_{.1}$, and $S_{|1.} = S + E_{1.}'C_{11}^{-1}E_{1.}$ and T stands for the matrix-T distribution.

The above result may be verified as follows. If $\begin{bmatrix} \Theta \\ \Sigma \end{bmatrix} \sim NW^{-1}(M, C, S, d)$ then $\Theta_{.1} \sim N(M_{.1}, C_{11}, \Sigma)$ given Σ and $\Sigma \sim W^{-1}(S, d)$, therefore $\Theta_{.1} \sim T(M_{.1}, C_{11}, S, d)$ and $\Sigma \sim W^{-1}(S_{|1.}, d + p_{1.})$ given $\Theta_{.1}$. Furthermore, $\Theta_{.2} \sim N(M_{2.|1.}, C_{22|1.}, \Sigma)$ given $\Theta_{.1}$ and Σ , thus $\begin{bmatrix} \Theta_{.2} \\ \Sigma \end{bmatrix} \sim NW^{-1}(M_{2.|1.}, C_{22|1.}, S_{|1.}, d + p_{1.})$ given $\Theta_{.1}$.

(d) Non-singular Matrix-normal inverted-Wishart

The random matrix $\begin{bmatrix} \Theta \\ \Sigma \end{bmatrix} \sim NW^{-1}(M, C, S, d)$ is referred to as non-singular if and only if both C and S are non-singular. In this case its probability density is, according to (A3.2.10) and (A3.2.15), given by,

$$p\left(\begin{bmatrix} \Theta \\ \Sigma \end{bmatrix}\right) = k(C, S, d) |\Sigma|^{-(\frac{1}{2}(d+p)+q)} \exp(-\frac{1}{2} \text{tr}(S + (\Theta - M)'C^{-1}(\Theta - M))\Sigma^{-1}), \quad (\text{A3.2.22})$$

where

$$k(C, S, d) = \frac{|S|^{\frac{1}{2}\nu}}{|C|^{\frac{1}{2}q} 2^{\frac{1}{2}(p+\nu)q} \pi^{\frac{1}{4}(q+2p-1)q} \prod_{i=1}^q ((\frac{1}{2}(\nu-i-1))!)}.$$

A3.2.4 Matrix-T Distribution.

Some properties of the matrix-T distribution, as defined in the previous section, follow.

(a) Marginal and conditional distributions:

From (A3.2.21) we immediately obtain the following result,

$$\begin{aligned} \Theta &\sim T(M, C, S, d) \quad \text{if and only if } \Theta_{1..} \sim T(M_{1..}, C_{11}, S, d) \quad \text{and} \\ \Theta_{2..} &\sim T(M_{2..|1..}, C_{22|1..}, S_{|1..}, d + p_{1..}) \quad \text{given } \Theta_{1..}, \end{aligned} \quad (\text{A3.2.23})$$

where the parameters are defined as in (A3.2.21). A similar result for subsets of rows is easily derived in view of (A3.2.26)

(b) Linear transformation:

From (A3.2.18) clearly,

$$\Theta \sim T(M, C, S, d) \quad \text{implies } H\Theta K + L \sim T(HMK + L, HCH', K'SK, d). \quad (\text{A3.2.24})$$

(c) Non-singular matrix-T distribution:

The random matrix $\Theta \sim T(M, C, S, d)$ is referred to as non-singular if and only if both C and S are non-singular. In this case its probability density is

$$p(\Theta) = \frac{p\left(\begin{bmatrix} \Theta \\ \Sigma \end{bmatrix}\right)}{p(\Sigma|\Theta)},$$

where $p\left(\begin{bmatrix} \Theta \\ \Sigma \end{bmatrix}\right)$ and $p(\Sigma|\Theta)$ are as in (A3.2.22) and the formula following (A3.2.19), i.e.

$$p(\Theta) = \left(\frac{|S|^{\frac{1}{2}\nu} \prod_{i=1}^q ((\frac{1}{2}(d+p+q-i-2))!) }{|C|^{\frac{1}{2}q} \pi^{\frac{1}{2}pq} \prod_{i=1}^q ((\frac{1}{2}(d+p-i-2))!) } \right) |S + (\Theta - M)'C^{-1}(\Theta - M)|^{-\frac{1}{2}(\nu+p)}. \quad (\text{A3.2.25})$$

Thus, Θ is distributed as the usual non-singular matrix-T distribution found, for example, in Zellner (1971, p. 396).

(d) Transpose:

$$\Theta \sim T(M, C, S, d) \quad \text{if and only if } \Theta' \sim T(M', S, C, d). \quad (\text{A3.2.26})$$

It is not difficult to show that (A3.2.26) is valid in the non-singular case; see for example Box and Tiao (1973, p. 442). The result can be verified for the general case as follows. Given any C, S there exist Q and R such that $C = QQ'$ and $S = R'R$. Thus, if $\Theta \sim T(M, C, S, d)$ there exists $\Xi \sim T(O, I, I)$ such that $M + Q\Xi R \sim T(M, C, S, d)$ and clearly $(M + Q\Xi R)' = M' + R'\Xi'Q' \sim T(M', S, C, d)$ as claimed.

(e) First and second moments:

$$\begin{aligned} \Theta \sim T(M, C, S, d) &\quad \text{implies} \\ \underset{\Theta}{\mathbb{E}} \operatorname{vec} \Theta = \operatorname{vec} M &\quad \text{and } \underset{\Theta}{\operatorname{VAR}} \operatorname{vec} \Theta = \frac{S}{d-2} \otimes C, \end{aligned} \quad (\text{A3.2.27a})$$

We can calculate these moments directly,

$$\underset{\Theta}{\mathbb{E}} \operatorname{vec} \Theta = \underset{\Sigma}{\mathbb{E}} \underset{\Theta|\Sigma}{\mathbb{E}} \operatorname{vec} \Theta = \underset{\Sigma}{\mathbb{E}} \operatorname{vec} M = \operatorname{vec} M$$

and

$$\underset{\Sigma}{\operatorname{VAR}} \operatorname{vec} \Theta = \underset{\Sigma}{\mathbb{E}} \underset{\Theta|\Sigma}{\operatorname{VAR}} \operatorname{vec} \Theta + \underset{\Sigma}{\operatorname{VAR}} \underset{\Theta|\Sigma}{\mathbb{E}} \operatorname{vec} \Theta = \underset{\Sigma}{\mathbb{E}} \Sigma \otimes C + \underset{\Sigma}{\operatorname{VAR}} M = \frac{S}{d-2} \otimes C.$$

As in the matrix-normal case, (A3.2.27a) can be rewritten as,

$$\begin{aligned} \Theta \sim T(M, C, S, d) &\quad \text{implies} \\ \underset{\Theta}{\mathbb{E}} \Theta = M, & \\ \underset{\Theta}{\mathbb{E}} (\Theta - M)(\Theta - M)' = C \operatorname{tr} \left(\frac{S}{d-2} \right) &\quad \text{and } \underset{\Theta}{\mathbb{E}} (\Theta - M)'(\Theta - M) = \frac{S}{d-2} \operatorname{tr}(C). \end{aligned} \quad (\text{A3.2.27b})$$

(f) Multivariate t distribution:

From the definition of the matrix-T distribution (Section A3.2.3 (c)) it is clear that, as in the matrix-normal case, the representation is not minimal. In other words, $T(M, C, S, d)$ and $T(M, \lambda C, \lambda^{-1}S, d)$ denote the same distribution. When the distribution concerns a column vector it is denoted as $t(\underline{m}, sC, d)$ and referred to as the multivariate t distribution.

CHAPTER 4

IMPLEMENTATION ASPECTS

In this chapter we discuss possible methods for the implementation of the DLMR updating recurrences. It is not surprising, in view of the relationship between the DLMR, DLM and state-space models, that the filter algorithms of the latter can be extended in order to cope with the first. In Section 4.1 we show how some well-known state-space filters can be generalized in order to cope with the DLMR. In 4.2 we present a new filter algorithm based on the ubiquitous sweep operator, and emphasis is put on its generality and simplicity. The necessary theory concerning the sweep operator is provided in Appendix A4.1.

4.1 IMPLEMENTATION VIA STATE-SPACE FILTERS.

The algorithms for the implementation of the updating recurrence (3.18) discussed in this section are generalizations of some well-known state-space filters: Kalman, Joseph, Square-root and Inverse-covariance. For the DLM with known σ^2 they are exactly equivalent. For this reason we retain the same terminology. The book by Maybeck (1979) is cited throughout this section and is referred as M1.

4.1.1 Kalman Filter.

The recurrence (3.18), where C_t is written as,

$$C_t = C_t^* - A_t X_t C_t^* = (I - A_t X_t) C_t^*, \quad (4.1)$$

is the Kalman filter version; compare with (M1, p. 209,217). \hat{E}_t is called innovation and A_t is referred to as the Kalman gain.

Although theoretically correct, the Kalman filter suffers from numerical difficulties, especially at the most critical stage of the observational update the form (4.1) fails to assure symmetry and positive definiteness for C_t (M1, p. 236, 237, 377) due to rounding errors.

4.1.2 Joseph Filter.

The Joseph form overcomes those problems associated with C_t by rewriting the measurement update formulas for M_t and C_t in (3.10c) as,

$$M_t = B_t M_t^* + A_t Y_t \quad \text{and} \quad C_t = B_t C_t^* B_t' + A_t V_t A_t', \quad (4.2)$$

where $B_t = I - A_t X_t$.

The algebraic equivalence is easily shown:

$$M_t = M_t^* + A_t(Y_t - X_t M_t^*) = (I - A_t X_t) M_t^* + A_t Y_t$$

and

$$\begin{aligned} C_t &= C_t - A_t \check{Y}_t A_t' + A_t(X_t C_t^* X_t' + V_t - \check{Y}_t) A_t' \\ &= C_t - A_t X_t C_t^* - C_t^* X_t A_t' + A_t X_t C_t^* X_t' A_t' + A_t V_t A_t' = (I - A_t X_t) C_t^* (I - A_t X_t)' + A_t V_t A_t'. \end{aligned}$$

The form (4.2) is better conditioned, and offers greater assurance of the symmetry and positive definiteness of C_t at the price of a greater number of computations (M1, p. 237). From a theoretical point of view (4.2) emphasizes the role of A_t as a generalized weight on the actual Y_t in comparison with past information.

4.1.3 Square-root Filter.

Although both the Kalman and Joseph filters can cope with most applications using a proper rescaling of variables and/or double precision, they are inherently numerically unstable. This situation is corrected by the implementation of a Square-root filter (M1, Chapter 7). The square-root filter works with the Cholesky square root of the system variances $V_t^c, W_t^c, C_{t-1}^c, S_{t-1}^c$ rather than with the variances themselves, and makes use of triangularization algorithms in order to obtain $C_t^{*c}, \check{Y}_t^c, C_t^c, S_t^c$. In so doing, the word-length required for a proper performance is obviously reduced, and simultaneously the symmetry and positive definiteness is completely assured.

The following notation is employed for the description of the filter. A^c stands for the Cholesky decomposition of A , i.e. $A = A^{c\prime} A^c$ and A^c is upper triangular, $R \leftarrow A$ denotes a triangularization algorithm, e.g. the modified Gram-Schmidt orthogonalization procedure or the Householder transformation (M1, p. 380,381), which yields R from A , where $A = QR$ is the QR decomposition of A , i.e. Q is orthogonal and R is upper triangular.

The filter algorithm is

Evolution:

$$M_t^* = G_t M_{t-1} \quad \text{and} \quad \begin{bmatrix} C_t^{*c} \\ O \end{bmatrix} \leftarrow \begin{bmatrix} W_t^c \\ C_{t-1}^c G_t' \end{bmatrix}. \quad (4.3a)$$

Prediction:

$$\hat{Y}_t = X_t M_t^* \quad \text{and} \quad \begin{bmatrix} \check{Y}_t^c & \tilde{A}_t' \\ O & C_t^c \end{bmatrix} \leftarrow \begin{bmatrix} V_t^c & O \\ C_t^{*c} X_t' & C_t^{*c} \end{bmatrix}. \quad (4.3b)$$

Posterior:

$$M_t = M_t^* + \tilde{A}_t \check{Y}_t^{c-1} \hat{E}_t, \quad \begin{bmatrix} S_t^c \\ O \end{bmatrix} \leftarrow \begin{bmatrix} S_{t-1}^c \\ \check{Y}_t^{c-1} \hat{E}_t \end{bmatrix}, \quad (4.3c)$$

and of course $d_t = d_{t-1} + \hat{E}_t$.

This recursion form may be verified as follows: The recurrence (4.3a) is correct since

$$C_t^* = \begin{bmatrix} C_t^{*c} \\ O \end{bmatrix}' \begin{bmatrix} C_t^{*c} \\ O \end{bmatrix} = \begin{bmatrix} W_t^c \\ C_{t-1}^c G_t' \end{bmatrix}' \begin{bmatrix} W_t^c \\ C_{t-1}^c G_t' \end{bmatrix} = W_t + G_t C_{t-1} G_t'.$$

The recurrence (4.3b) is valid since

$$\begin{aligned} \begin{bmatrix} \check{Y}_t & \check{Y}_t^c \tilde{A}_t' \\ \tilde{A}_t \check{Y}_t & \tilde{A}_t \tilde{A}_t' + C_t \end{bmatrix} &= \begin{bmatrix} \check{Y}_t^c & \tilde{A}_t' \\ O & C_t^c \end{bmatrix}' \begin{bmatrix} \check{Y}_t^c & \tilde{A}_t' \\ O & C_t^c \end{bmatrix} \\ &= \begin{bmatrix} V_t^c & O \\ C_t^{*c} X_t' & C_t^{*c} \end{bmatrix}' \begin{bmatrix} V_t^c & O \\ C_t^{*c} X_t' & C_t^{*c} \end{bmatrix} = \begin{bmatrix} V_t + X_t C_t^* X_t' & X_t C_t^* \\ C_t^{*c} X_t' & C_t^{*c} \end{bmatrix}, \end{aligned}$$

i.e. $\check{Y}_t = V_t + C_t^* X_t'$, $\tilde{A}_t = C_t^* X_t' \check{Y}_t^{c-1}$ and $C_t = C_t^* - C_t^* X_t' \check{Y}_t^{-1} X_t C_t^*$. Finally, (4.3c) means that $M_t = M_t^* + A_t \hat{E}_t$ since $\tilde{A}_t \check{Y}_t^{c-1} t' = C_t^* X_t' \check{Y}_t^{c-1} \check{Y}_t^{c-1} t' = C_t^* X_t' \check{Y}_t^{-1}$ and

$$\begin{aligned} S_t &= \begin{bmatrix} S_t^c \\ O \end{bmatrix}' \begin{bmatrix} S_t^c \\ O \end{bmatrix} = \begin{bmatrix} S_t^c \\ \check{Y}_t^{c-1} \hat{E}_t \end{bmatrix}' \begin{bmatrix} S_t^c \\ \check{Y}_t^{c-1} \hat{E}_t \end{bmatrix} \\ &= S_{t-1} + \hat{E}_t \check{Y}_t^{c-1} \check{Y}_t^{c-1} \hat{E}_t = S_{t-1} + \hat{E}_t \check{Y}_t^{-1} \hat{E}_t. \end{aligned}$$

It is important to note how the requirements of the procedure can be satisfied: the inverse of \check{Y}_t^c required in (4.3c) can be rapidly obtained due to the triangular form of \check{Y}_t^c . The Cholesky square-roots corresponding to C and S have to be computed only once since the filter updates them for the following period and, in general, V_t^c and W_t^c have to be obtained for each period. For the computation of C_o^c , S_o^c , V_t^c and W_t^c a standard Cholesky algorithm (M1, p. 371) may be employed; see also A4.1.3.

4.1.4 Inverse-covariance Filter.

Another recurrence for the measurement update, in terms of the inverses of C_t and V_t , is known as the inverse-covariance form (M1, p. 238-242). The recurrence of this filter

$$\begin{aligned} M_t &= C_t(C_t^{*-1} M_t^* + X_t' V_t^{-1} Y_t), \quad C_t^{-1} = C_t^{*-1} + X_t' V_t^{-1} X_t \quad \text{and} \\ S_t &= S_{t-1} + M_t^{*'} C_t^{*-1} M_t^* + Y_t' V_t^{-1} Y_t - M_t' C_t^{-1} M_t. \end{aligned} \tag{4.4}$$

The correctness of (4.4) may be established by means of the binomial inverse theorem (A4.1.6) after a good deal of algebra. A direct derivation through the sweep operator is given in the next section. This form also resembles, in fact generalizes, some familiar formulas which appear in the literature of Bayesian analysis of linear models, e.g. De Groot (1970, p. 251-252) and Zellner (1971, p. 234-235). Often the variance inverses are referred to as precisions. In this context, the form (4.4) supports the terminology "weighted", depending on V_t , used in Chapter 3; see also the comment after recurrence (4.2). Moreover, for the particular vague prior $NW^{-1}(M_t^*, C_t^*, S_{t-1}, d_{t-1})$ with $C_t^* = \epsilon^{-1} I$, $S_{t-1} = \epsilon I$ (and say $M_t^* = O$, $d = \epsilon$), the form (4.4) leads to,

$$M_t = C_t X_t' V_t^{-1} Y_t, \quad C_t = (X_t' V_t^{-1} X_t)^{-1} \quad \text{and} \quad S_t = (Y_t - X_t M_t)' V_t (Y_t - X_t M_t), \tag{4.5}$$

as $\epsilon \rightarrow 0^+$.

These equations establish a relationship between the DLMR recurrence and several results of the multivariate non-Bayesian static linear models. For the standard multivariate regression, M_t , $\Sigma \otimes C_t$ and S_t are respectively the maximum likelihood estimator of Θ_t , the variance of $\text{vec } \Theta_t$ and the residual sum of squares. For the weighted generalized multivariate regression, M_t , ΣC_t and S_t are the Aitken estimator of the regressor parameters, its variance and the weighted residual sum of squares; compare with Press (1982, p. 229-247). A non-Bayesian counterpart of the DWMR has been independently developed by Harvey (1986). He shows that for the non-weighted steady model ($V_t = 1$, $W_t = w$, $X_t = 1$, $G_t = 1$), in the absence of any prior information (whatever that means in a non-Bayesian context!), the minimum mean square estimator for $\Theta_t = \underline{\theta}_t'$ is $M_t^* = \underline{m}_t^{*'} \underline{m}_t$ and its mean square error is ΣC_t^* . In

addition, the maximum likelihood estimator of Σ is $\frac{S_t}{d_t - 1}$. The starting values proposed by him at time $t = 2$ are $m_2^* = y_1 = Y'_1, C_2^* = 1 + w, S_1 = 0$ and $d_1 = 0$. He also suggests that the DWMR can be handled in an analogous way, and in particular, that the implementation can be achieved by applying the univariate Kalman filter to each series in turn. Although this latter suggestion, in principle, is correct as it is mentioned in Section 3.2, this form of implementation is very inefficient as opposed to the multivariate Kalman filter and/or to the sweep operator recurrence; see Highfield (1984) and Quintana (1985).

From the numerical point of view (4.4) presents some advantages for those cases in which C_t^{-1} and/or V_t^{-1} are very close to O , for example, when there is great uncertainty at the beginning of the analysis of the time series.

4.2 IMPLEMENTATION VIA THE SWEEP OPERATOR.

All filters considered so far assume implicitly the existence of \check{Y}_t^{-1} . Not surprisingly, difficulties arise when \check{Y}_t is singular. Although it is possible to use the Kalman filter formulas replacing \check{Y}_t^{-1} by the pseudo-inverse of \check{Y}_t , an algorithm for obtaining such pseudo-inverse is necessary.

Fortunately, there is an direct and simple alternative for overcoming these problems which is based on the sweep operator (Quintana, 1985). The major advantages of this implementation are: it can be easily programmed; it yields the desired result either for \check{Y}_t singular or non-singular; and in addition it is suitable for finding the conditional distribution of a subset of columns (rows) of Y_t given another subset of columns (rows). On the other hand, as the Kalman filter, this implementation fails to assure positive definiteness, but again it can cope with most applications using a proper rescaling of variables and/or double precision. The relevant theory about the sweep operator is developed in Appendix 4.1.

A very important related topic is the use of the sweep operator for the implementation of a dynamic version of the step-wise regression. This is examined in Section 6.3.

4.2.1 Updating Recurrence given Σ .

We start our discussion by looking at the particular case of the DLMR which can be easily reduced to the conventional DLM. Recently Chen (1985, p. 222) has pointed out that the Kalman filter formulas remain valid if \check{Y}_t^{-1} is replaced by the pseudo-inverse of \check{Y}_t . Not so recently, Dempster (1969, p. 277-278) gave the solution for the essential problem of obtaining the conditional distribution for a general, possibly singular, multivariate normal distribution and he outlined a practical procedure based on sweep operations. Here, an updated recurrence that follows the latter approach is presented.

Formula (3.8) obviously can be rewritten as,

$$\begin{bmatrix} Y_t \\ \Theta_t \end{bmatrix} \sim N \left(\begin{bmatrix} \check{Y}_t \\ M_t^* \end{bmatrix}, \begin{bmatrix} \check{Y}_t & X_t C_t^* \\ C_t^* X_t' & C_t^* \end{bmatrix}, \Sigma \right). \quad (4.6)$$

Thus, results (A3.2.8) and (A4.1.13) imply that M_t and C_t in (3.7c) can be obtained via the sweep operation,

$$\begin{bmatrix} \check{Y}_t & X_t C_t^* & -\hat{E}_t \\ C_t^* X_t' & C_t^* & M_t^* \end{bmatrix} \rightarrow \begin{bmatrix} \check{Y}_t^{-1} & A_t' & -\check{Y}_t^{-1} \hat{E}_t \\ -A_t & C_t & M_t \end{bmatrix}, \quad (4.7)$$

provided that \check{Y}_t is non-singular. Better still, (A4.1.16) shows that the subsweep operation,

$$\begin{bmatrix} \check{Y}_t & X_t C_t^* & -\hat{E}_t \\ C_t^* X_t' & C_t^* & M_t^* \end{bmatrix} \xrightarrow{\check{Y}_t} [C_t \quad M_t], \quad (4.8)$$

always yields the updated hyper-parameters regardless of whether \check{Y}_t (or any other variance) is singular or non-singular. Furthermore, (4.8) is significantly more efficient than (4.7) when r is large, in particular for a conventional multivariate DLM.

4.2.2 Updating Recurrence when Σ is Unknown.

Proceeding in an analogous form, it is not difficult to see that (4.6), (3.6a), (A3.2.17), (A3.2.21) together with (A4.1.17) and (A4.1.18) imply that M_t, C_t and S_t can be obtained either applying the sweep operation,

$$\begin{bmatrix} \check{Y}_t & X_t C_t^* & -\hat{E}_t \\ C_t^* X_t' & C_t^* & M_t^* \\ \hat{E}_t' & -M_t^{**} & S_{t-1} \end{bmatrix} \rightarrow \begin{bmatrix} \check{Y}_t^{-1} & A_t' & -\check{Y}_t^{-1} \hat{E}_t \\ -A_t & C_t & M_t \\ -\hat{E}_t \check{Y}_t^{-1} & -M_t' & S_t \end{bmatrix}, \quad (4.9)$$

or more generally via the subsweep operation,

$$\begin{bmatrix} \check{Y}_t & X_t C_t^* & -\hat{E}_t \\ C_t^* X_t' & C_t^* & M_t^* \\ \hat{E}_t' & -M_t^{**} & S_{t-1} \end{bmatrix} \xrightarrow{\check{Y}_t} \begin{bmatrix} C_t & M_t \\ -M_t' & S_t \end{bmatrix}. \quad (4.10)$$

It is remarkable that the simple transformation (4.10) provides the updating solution for the DLMR model regardless of whether the matrices V_t, W_t, C_t and/or $\Sigma (S_t)$ are singular or non-singular. Furthermore, even the sufficient statistics for those non-Bayesian static models described in Subsection 4.1.4 still can be obtained, in practice, from (4.10) by setting ϵ in (4.5) equal to small positive value, for instance, the square root of the particular machine precision.

Also apparent, because of its sequential nature, is the ease of (4.10) for handling missing rows of observations. This is particularly important for multivariate DLM's.

4.2.3 Contemporaneous Conditional Predictive Distributions.

A major feature of multivariate time series models is that they provide a means for updating our predictive probabilities for a subset of observations when another subset of contemporaneous observations is given. An important application of these updated predictive distributions is pooling external forecasts.

In the description of the use of the subsweep operator for finding the predictive distributions of subsets of columns(rows) of Y_t given another subset of columns(rows), we follow the notation of Appendix A3.2, i.e.

$$Y_t = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}_t = \begin{bmatrix} Y_{1.} \\ Y_{2.} \end{bmatrix}_t = [Y_{1.}, Y_{2.}]_t, \quad \text{etc.}$$

In accordance with (3.10b), (A3.2.23) and (A4.1.18) the predictive distribution of $Y_{2.t}$ given $Y_{1.t}$ is,

$$Y_{2.t}|Y_{1.t} \sim T(\hat{Y}_{2.|1.t}, \check{Y}_{22|1.t}, S_{|1.(t-1)}, d_{|1.(t-1)}), \quad (4.11)$$

where the parameters can be obtained by performing the subsweep operation,

$$\begin{bmatrix} \check{Y}_{11t} & \check{Y}_{12t} & -\hat{E}_{1,t} \\ \check{Y}_{21t} & \check{Y}_{22t} & \hat{Y}_{2,t} \\ \hat{E}'_{1,t} & -\hat{Y}'_{2,t} & S_{t-1} \end{bmatrix} \xrightarrow{\check{Y}_{11t}} \begin{bmatrix} \check{Y}_{22|1,t} & \hat{Y}_{2,|1,t} \\ -\hat{Y}'_{2,|1,t} & S_{|1,(t-1)} \end{bmatrix}; \quad (4.12)$$

see also the relevant comment following formula (A4.1.18) concerning the shape parameter $d_{|1,(t-1)}$.

The procedure when Σ is known merely deletes the bottom row of the partitioned matrices in (4.12), i.e.

$$Y_{2,t}|Y_{1,t} \sim N(\hat{Y}_{2,|1,t}, \check{Y}_{22|1,t}, \Sigma), \quad (4.13)$$

where the parameters are given by,

$$\begin{bmatrix} \check{Y}_{11t} & \check{Y}_{12t} & -\hat{E}_{1,t} \\ \check{Y}_{21t} & \check{Y}_{22t} & \hat{Y}_{2,t} \end{bmatrix} \xrightarrow{\check{Y}_{11t}} \begin{bmatrix} \check{Y}_{22|1,t} & \hat{Y}_{2,|1,t} \end{bmatrix}. \quad (4.14)$$

Similarly, the distribution of $Y_{2t}|Y_{1t}$ is,

$$Y_{2t}|Y_{1t} \sim T(\hat{Y}_{2,|1,t}, \check{Y}_t, S_{22|1(t-1)}, d_{|1(t-1)}), \quad (4.15)$$

where the parameters are given by,

$$\begin{bmatrix} \check{Y}_t & -\hat{E}_{1,t} & \hat{Y}_{2,t} \\ \hat{E}'_{1,t} & S_{11(t-1)} & S_{12(t-1)} \\ -\hat{Y}'_{2,t} & S_{21(t-1)} & S_{22(t-1)} \end{bmatrix} \xrightarrow{S_{11(t-1)}} \begin{bmatrix} \check{Y}_{1,1t} & \hat{Y}_{2,1t} \\ -\hat{Y}_{2,1t} & S_{22|1(t-1)} \end{bmatrix}. \quad (4.16)$$

If Σ is known then $Y_{2t}|Y_{1t} \sim N(\hat{Y}_{2,|1,t}, \check{Y}_t, \Sigma_{22|1})$ where the parameters are obtained as in (4.16) substituting S_{t-1} by Σ .

Applying sequentially (4.12) and (4.16) more conditional distributions can be obtained, for instance, the distribution of $Y_{12t}|Y_{11t}, Y_{21t}, Y_{22t}$ follows conditioning first on $[Y_{21t}, Y_{22t}]$ and then on Y_{11t} . Of course, with obvious modifications the procedure may be employed when Σ is known; see also O'Hagan (1972).

Furthermore, if the subsweep operation (4.12) and/or (4.16) is applied to the full matrix appearing in the left-hand side of (4.10) then the distribution of the corresponding model parameters are obtained in addition to the updated predictive distribution; see Section 6.3.

4.2.4 Derivation of the Inverse Covariance Filter Recurrence.

The sweep operator not only provides a versatile and powerful pivoting scheme useful for implementing purposes, but in addition it is an elegant tool which may be employed for getting theoretical results. In order to prove our point we present a neat derivation of the inverse covariance filter recurrence formulas (4.4) which parallels our derivation of the binomial inverse theorem formulas (A4.1.6) given in Appendix A4.

The outline of the proof is, having in mind the order-independence and reflexivity of the sweep operator, very simple: by pivoting the left-hand side matrix in (4.9) on (the diagonal elements of) C_t^* ,

then pivoting on the corresponding matrix of \check{Y}_t , and pivoting back on the corresponding matrix of C_t^* the recurrence (4.4) equivalent to (3.10c) is obtained.

The detailed derivation proceeds as follows. The first operation, pivoting on C_t^* , is,

$$\begin{bmatrix} \check{Y}_t & X_t C_t^* & -\hat{E}_t \\ C_t^* X_t' & C_t^* & M_t^* \\ \hat{E}_t' & -M_t^{**'} & S_{t-1} \end{bmatrix} \rightarrow \begin{bmatrix} \check{Y}_t - X_t C_t^* X_t' & -X_t & -(\hat{E}_t + X_t M_t^*) \\ X_t' & C_t^{*-1} & C_t^{*-1} M_t^* \\ \hat{E}_t' + M_t^{**'} X_t' & M_t^{**'} C_t^{*-1} & S_{t-1} + M_t^{**'} C_t^{*-1} M_t^* \end{bmatrix}, \quad (4.18)$$

but according to the notation introduced in (3.7) the right-hand side of (4.18) can be rewritten as the left-hand side of (4.19), which is the second operation, pivoting on V_t ,

$$\begin{aligned} & \begin{bmatrix} V_t & -X_t & -Y_t \\ X_t' & C_t^{*-1} & C_t^{*-1} M_t^* \\ Y_t' & M_t^{**'} C_t^{*-1} & S_{t-1} + M_t^{**'} C_t^{*-1} M_t^* \end{bmatrix} \\ & \rightarrow \begin{bmatrix} V_t^{-1} & -V_t^{-1} X_t & -V_t^{-1} Y_t \\ -X_t' V_t^{-1} & C_t^* + X_t' V_t^{-1} X_t & C_t^{*-1} M_t^* + X_t' V_t^{-1} Y_t \\ -Y_t' V_t^{-1} & M_t^{**'} C_t^{*-1} + Y_t' V_t^{-1} X_t & S_{t-1} + M_t^{**'} C_t^{*-1} M_t^* + Y_t' V_t^{-1} Y_t \end{bmatrix}. \end{aligned} \quad (4.19)$$

Finally, pivoting back the right-hand side of (4.19) on $C_t^{*-1} + X_t' V_t^{-1} X_t$ yields a matrix which is necessarily equal to the right-hand side of (4.9) and recurrence (4.4) follows immediately. In addition, we obtain the following formulas as a side-product,

$$\begin{aligned} A_t &= C_t X_t' V_t^{-1}, \quad \check{Y}_t^{-1} \hat{E}_t = V_t^{-1} (Y_t - X_t M_t) \quad \text{and} \\ \check{Y}_t^{-1} &= V_t^{-1} - A_t' C_t^{-1} A_t. \end{aligned} \quad (4.20)$$

Following similar pivoting patterns other equivalent recurrences can be obtained, in fact, the number of possible ways grows rapidly depending on r, p and q . However, among those the direct operation (4.9), intrinsically related to recurrence (3.10c), is the most economical in the sense that no double pivoting (pivoting back) is involved in the process.

APPENDIX A4.1.

THE SWEEP OPERATOR.

The sweep operator is a very useful tool which has many applications in statistics. A good review may be found in Goodnight (1979) and implementation aspects in Clarke (1981). The sweep operator has also applications in quadratic programming where it is known as principal pivoting (Berman and Plemmons, 1979; Quintana, O'Reilly and Gómez, 1986).

For completeness, several results of the sweep operator relevant for applications to the DLMR are derived.

A4.1.1 Linear Equations and Pivotal Operations.

Consider two sets of variables \underline{w} and \underline{z} which satisfy the linear equation,

$$\underline{w} = -A\underline{z}. \quad (\text{A4.4.1})$$

This equation can be represented in tableau form as,

$$\underline{w} \begin{array}{c} \underline{z}' \\ \hline A, \end{array} \quad (\text{A4.1.2})$$

where \underline{w} is referred to as basic variables and \underline{z} as non-basic variables.

The question is: which is the new tableau when a subset of basic variables, say \underline{w}_1 , is swapped for the corresponding non-basic variables \underline{z}_1 , where $\underline{w} = \begin{bmatrix} \underline{w}_1 \\ \underline{w}_2 \end{bmatrix}$ and $\underline{z} = \begin{bmatrix} \underline{z}_1 \\ \underline{z}_2 \end{bmatrix}$? The equation (A4.1.1) is equivalent to,

$$\underline{w}_1 = -A_{11}\underline{z}_1 - A_{12}\underline{z}_2, \quad (\text{A4.1.3a})$$

$$\underline{w}_2 = -A_{21}\underline{z}_1 - A_{22}\underline{z}_2. \quad (\text{A4.1.3b})$$

Obtaining \underline{z}_1 from (A4.1.3a) and substituting in (A4.1.3b) gives

$$\underline{z}_1 = -A_{11}^{-1}\underline{w}_1 - A_{11}^{-1}A_{12}\underline{z}_2, \quad (\text{A4.1.4a})$$

$$\underline{w}_2 = A_{21}A_{11}^{-1}\underline{w}_1 - (A_{22} - A_{21}A_{11}^{-1}A_{12})\underline{z}_2. \quad (\text{A4.1.4b})$$

In terms of the tableau associated with (A4.1.3) and (A4.1.4), the process is represented as

$$\begin{array}{l} \underline{w}_1 \begin{bmatrix} \underline{z}'_1 & \underline{z}'_2 \\ \hline A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \rightarrow \underline{w}_1 \begin{bmatrix} \underline{z}'_1 & \underline{w}'_1 & \underline{z}'_2 \\ \hline A_{11}^{-1} & A_{11}^{-1}A_{12} & A_{11}^{-1}A_{12} \\ -A_{21}A_{11}^{-1} & A_{22} - A_{21}A_{11}^{-1}A_{12} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix} \\ \underline{w}_2 \end{array} \quad (\text{A4.1.5})$$

The right-hand matrix between brackets is referred to as the result of pivoting (or sweeping) $\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ on A_{11} .

The sweep operation (A4.1.5) is a reflexive operation in the sense that pivoting

$$\begin{bmatrix} A_{11}^{-1} & A_{11}^{-1}A_{12} \\ -A_{21}A_{11}^{-1} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}$$

on A_{11}^{-1} reproduces A since this is equivalent to swapping-back the swapped variables. It is also clear that it is possible to swap the variables sequentially i.e. sweeping

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \quad \text{on } \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

is equivalent to sweeping A on A_{11} first and then sweeping the result on $A_{22} - A_{21}A_{11}^{-1}A_{12}$ assuming, of course, that the involved inverses exist. Moreover, the sweeping order is irrelevant in the sense that sweeping A on A_{22}^{-1} and then the result on $A_{11} - A_{12}A_{22}^{-1}A_{21}$ yields the same outcome.

For example, consider the square matrix $\begin{bmatrix} A & -B \\ C & D \end{bmatrix}$. The inverse can be obtained as follows,

$$\begin{aligned} \frac{w_1}{w_2} \begin{bmatrix} z'_1 & z'_2 \\ A & -B \\ C & D \end{bmatrix} &\rightarrow \frac{z_1}{w_2} \begin{bmatrix} w'_1 & z'_2 \\ A^{-1} & -A^{-1}B \\ -CA^{-1} & D + CA^{-1}B \end{bmatrix} \\ &\rightarrow \frac{z_1}{z_2} \begin{bmatrix} w'_1 & w'_2 \\ A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1} & A^{-1}B(D + CA^{-1}B)^{-1} \\ -(D + CA^{-1}B)^{-1}CA^{-1} & (D + CA^{-1}B)^{-1} \end{bmatrix} \end{aligned}$$

or equivalent as

$$\begin{aligned} \frac{w_1}{w_2} \begin{bmatrix} z'_1 & z'_2 \\ A & -B \\ C & D \end{bmatrix} &\rightarrow \frac{w_1}{z_2} \begin{bmatrix} z'_1 & w'_2 \\ A + BD^{-1}C & BD^{-1} \\ D^{-1}C & D^{-1} \end{bmatrix} \\ &\rightarrow \frac{z_1}{z_2} \begin{bmatrix} w'_1 & w'_2 \\ (A + BD^{-1}C)^{-1} & -(A + BD^{-1}C)^{-1}BD^{-1} \\ D^{-1}C(A + BD^{-1}C)^{-1} & D^{-1} - D^{-1}C(A + BD^{-1}C)^{-1}BD^{-1} \end{bmatrix} \end{aligned}$$

Therefore, the following symmetric identities are obtained,

$$(A + BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}, \quad (\text{A4.1.6a})$$

$$-(A + BD^{-1}C)^{-1}BD^{-1} = A^{-1}B(D + CA^{-1}B)^{-1}, \quad (\text{A4.1.6b})$$

$$D^{-1}C(A + BD^{-1}C)^{-1} = -(D + CA^{-1}B)^{-1}CA^{-1}, \quad (\text{A4.1.6c})$$

$$D^{-1} - D^{-1}C(A + BD^{-1}C)^{-1}BD^{-1} = (D + CA^{-1}B)^{-1}, \quad (\text{A4.1.6d})$$

where the indicated inverses are assumed to exist. The symmetric identities (A4.1.6a) and (A4.1.6d) are known as the binomial inverse theorem and also as the matrix inversion lemma ; see for example Press (1982, p. 23) and Anderson and Moore (1979, p. 138).

A4.1.2 Multivariate Regression.

Let $[X, Y]$ be data from the following linear model,

$$Y = XB + E. \quad (\text{A4.1.7})$$

Then the least-squares solution $\hat{B} = (X'X)^{-1}X'Y$ can be obtained by performing a sweep operation onto the cross-product data matrix $[X, Y]'[X, Y]$,

$$\begin{bmatrix} X'X & X'Y \\ Y'X & Y'Y \end{bmatrix} \rightarrow \begin{bmatrix} (X'X)^{-1} & (X'X)^{-1}X'Y \\ -Y'X(X'X)^{-1} & Y'Y - Y'X(X'X)^{-1}X'Y \end{bmatrix} = \begin{bmatrix} (X'X)^{-1} & \hat{B} \\ -\hat{B}' & \hat{E}'\hat{E} \end{bmatrix}, \quad (\text{A4.1.7})$$

where $\hat{E} = Y - X\hat{B}$.

A careful look at (A4.1.7) reveals that the transformation always can be carried out sequentially regardless of the rank of X , by skipping the null pivots since a null pivot means that its corresponding independent variable is an exact linear combination of the independent variables already included in the regression, and it is therefore redundant. Furthermore, if $X'X$ is positive definite then null pivots never arise.

In view of the reflexive property of the sweep operator the scheme (A4.1.7) has an immediate application in stepwise-regression, in fact, this was the original application in Efroymson (1960). An obvious criticism is that the machine precision must be great enough for storing the largest element of the cross-product data matrix.

A4.1.3 Cholesky Decomposition.

The Cholesky decomposition of a non-negative definite symmetric matrix S is

$$S = R'R, \quad (\text{A4.1.8})$$

where R is a upper triangular matrix.

The sweep operator provides a simple method for finding R . Let

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} R_{11} & R_{12} \\ O & R_{22} \end{bmatrix},$$

then a little algebra shows that (A4.1.8) implies,

$$S_{11} = R'_{11}R_{11}, \quad S_{12} = R'_{11}R_{12} \quad \text{and} \quad (\text{A4.1.9a})$$

$$S_{22} - S_{21}S_{11}^{-1}S_{12} = R'_{22}R_{22}. \quad (\text{A4.1.9b})$$

Therefore, taking S_{11} to be a scalar, the first row of R can be obtained easily from (A4.1.9a) (if $A_{11} = O$ then we can take $R_{11} = O$ and say $R_{12} = O$). The Schur complement $S_{22} - S_{21}S_{11}^{-1}S_{12}$ can be calculated using (A4.1.5). Thus we have reduced the problem to finding the Cholesky decompositions of the Schur complement of S_{11} relative to S , and of course, we can perform the same procedure over and over again as long as necessary.

Let S be a positive semidefinite symmetric matrix, then from (A4.1.8) and (A4.1.9b) it is clear that a sequential application of the sweep operator will end, reordering if necessary, as

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \rightarrow \begin{bmatrix} S_{11}^{-1} & S_{11}^{-1}S_{12} \\ -S_{21}S_{11}^{-1} & O \end{bmatrix}. \quad (\text{A4.1.10})$$

Note that the rank of S is the dimension of S_{11} . Moreover, pivoting back (A4.1.10) on S_{11}^{-1} shows that S can be written as,

$$S = [I, S_{11}^{-1}S_{12}]' S_{11} [I, S_{11}^{-1}S_{12}]. \quad (\text{A4.1.11})$$

A4.1.4 Implementation.

As mentioned before the transformation (A4.1.5) can be carried on by sweeping sequentially (the order is irrelevant) on the diagonal elements of A_{11} . It is also clear from (A4.1.5) that the following algorithm sweeps A on a_{kk} overwriting A :

1. $a_{kk} \leftarrow \frac{1}{a_{kk}}$.
2. $a_{ij} \leftarrow a_{ij} - a_{ik}a_{kk}a_{kj}$, for $i, j \neq k$.
3. $a_{kj} \leftarrow a_{kj}a_{kk}$, for $j \neq k$.
4. $a_{ik} \leftarrow -a_{ik}a_{kk}$, for $i \neq k$.

Thus the implementation of the sweep operator is very simple provided that A_{11} is positive definite (see comment after A4.1.7); for our purposes this is enough.

When A is square it is possible to take advantage of the symmetry apparent in (A4.1.5) in order to save memory requirements by operating only on the upper triangular matrix; see Goodnight (1979) and also Clarke (1982).

A 4.1.5 Matrix-variate Distributions.

The sweep operator can be applied neatly for finding normal, normal inverted-Wishart, and matrix-T conditional distributions. First, we establish a key separation principle for the sweep operator.

Consider the sweep operation A on A_{11} ,

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \rightarrow \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix}. \quad (\text{A4.1.12})$$

Thus, applying the formula (A4.1.5) it can be readily seen that B is given by,

$$\begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix} = \begin{bmatrix} A_{11}^{-1} & A_{11}^{-1}A_{12} & A_{11}^{-1}A_{13} \\ -A_{21}A_{11}^{-1} & A_{22} - A_{21}A_{11}^{-1}A_{12} & A_{23} - A_{21}A_{11}^{-1}A_{13} \\ -A_{31}A_{11}^{-1} & A_{32} - A_{31}A_{11}^{-1}A_{12} & A_{33} - A_{31}A_{11}^{-1}A_{13} \end{bmatrix}.$$

It is now apparent that the submatrices in B depend on the submatrices in A only through the corners of the parallelogram defined by their corresponding matrices in A and the pivoting matrix, i.e. B_{ij} depends on A only through A_{11}, A_{i1}, A_{1j} and A_{ij} for $i, j = 1, 2, 3$. Furthermore, the transformation (A4.1.12) can be seen as several separated sweep operations on A_{11}

$$A_{11} \rightarrow B_{11}, [A_{11}, A_{12}] \rightarrow [B_{11}, B_{12}], [A_{11}, A_{13}] \rightarrow [B_{11}, B_{13}], \dots, \\ \begin{bmatrix} A_{11} & A_{12} \\ A_{31} & A_{32} \end{bmatrix} \rightarrow \begin{bmatrix} B_{11} & B_{12} \\ B_{31} & B_{32} \end{bmatrix}, \dots, \begin{bmatrix} A_{11} & A_{13} \\ A_{31} & A_{33} \end{bmatrix} \rightarrow \begin{bmatrix} B_{11} & B_{13} \\ B_{31} & B_{33} \end{bmatrix}.$$

The extension by induction for more refined partitions can be easily appreciated. For instance, the algorithm given in (A4.1.4) illustrates the principle when the submatrices of A are the scalar elements. The principle also implies that multiplying the sequential pivots, during the application of the algorithm, yields the determinant of the pivoting matrix A_{11} . This can be shown by induction, using a well known formula (Noble and Daniel, 1977, p. 210): $|A_{11}| = |C_{11}| |C_{22} - C_{21}C_{11}^{-1}C_{12}|$ where $A_{11} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$.

Conditional Matrix-normal Distribution.

It is evident, keeping in mind the separation principle, that the parameters of the matrix-normal conditional distribution (A3.2.8) can be easily obtained through the sweep operation

$$\begin{bmatrix} C_{11} & C_{12} & -E_{1.} \\ C_{21} & C_{22} & M_{2.} \end{bmatrix} \rightarrow \begin{bmatrix} C_{11}^{-1} & C_{11}^{-1}C_{12} & -C_{11}^{-1}E_{1.} \\ -C_{21}C_{11}^{-1} & C_{22|1.} & M_{2|1.} \end{bmatrix}. \quad (\text{A4.1.13})$$

This transformation generalizes a previous use of the sweep operator by Dempster (1982) in order to obtain conditional variances in multivariate normal distributions.

A closer examination of (A4.1.13) reveals that we are concerned only with the Schur complement of C_{11} relative to

$$\begin{bmatrix} C_{11} & C_{12} & -E_{1.} \\ C_{21} & C_{22} & M_{2.} \end{bmatrix}, \quad \text{i.e. } [C_{22|1.}, M_{2|1.}]$$

Therefore, the sequential sweeping algorithm described in A4.1.4, can be modified in order to obtain $[C_{22|1.}, M_{2|1.}]$ in a more efficient way by discarding sequentially the rows and columns associated with each pivot (recall again the separation principle). Moreover, the algorithm can be interpreted as finding the conditional distribution of $\Theta_2.$ given $\Theta_1.$ by conditioning sequentially on the rows of $\Theta_1..$ To see this, suppose that we want to obtain the (parameters of the) distribution of $\Theta_3.$ given $\Theta_1.$ and $\Theta_2.$ where

$$\Theta = \begin{bmatrix} \Theta_1. \\ \Theta_2. \\ \Theta_3. \end{bmatrix} \sim N(M, C, \Sigma).$$

First we condition $\Theta_2.$ and $\Theta_3.$ on $\Theta_1.$ by performing the sweep operation (in an obvious notation),

$$\begin{bmatrix} C_{11} & C_{12} & C_{13} & -E_{1.} \\ C_{21} & C_{22} & C_{23} & -E_{2.} \\ C_{31} & C_{32} & C_{33} & M_{3.} \end{bmatrix} \rightarrow \begin{bmatrix} C_{11}^{-1} & C_{11}^{-1}C_{12} & C_{11}^{-1}C_{13} & C_{11}^{-1}E_{1.} \\ -C_{21}C_{11}^{-1} & C_{22|1.} & C_{23|1.} & -E_{2|1.} \\ -C_{31}C_{11}^{-1} & C_{32|1.} & C_{33|1.} & M_{3|1.} \end{bmatrix}. \quad (\text{A4.1.14})$$

Then the parameters $M_{3|1.,2.}, C_{33|1.,2.}$ in $\Theta_3.| \Theta_1., \Theta_2. \sim N(M_{3|1.,2.}, C_{33|1.,2.}, \Sigma)$ can be obtained by sweeping $\begin{bmatrix} C_{22|1.} & C_{23|1.} & -E_{2|1.} \\ C_{32|1.} & C_{33|1.} & M_{3|1.} \end{bmatrix}$ on $C_{22|1.}$ i.e. conditioning $\Theta_3.| \Theta_1.$ on $\Theta_2..$ Thus, the modified algorithm described above is simply the consequence of refining the partitions to the limit.

Not surprisingly, in view of its sequential character, the algorithm can be further modified in order to deal with positive semidefinite variances according to the discussion after formula (A3.2.8). Let us assume that $\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$ is positive semidefinite. Then, without loss of generality, reordering if necessary as in (A4.1.10), $M_{3|2.,1.}$ and $C_{33|1.,2.}$ are obtained directly from (A4.1.14), i.e. $M_{3|2.,1.} = M_{3|1.}$ and $C_{33|1.,2.} = C_{33|1.}$ since according to (A4.1.10) $C_{22|1.} = O$ and therefore $\Theta_2. = M_{2|1.} = M_2. + C_{21}C_{11}^{-1}E_{1.} = M_2. + C_{21}C_{11}^{-1}(\Theta_1. - M_1.)$ given $\Theta_1..$ In other words, $\Theta_2.$ is a linear variant of $\Theta_1..$ and provides no further information given $\Theta_1..$ Notice that in practice no reordering is necessary. When a dependence is encountered the algorithm handles it merely by skipping the offending null pivot. The operator induced by this general algorithm is referred to as the subsweep operator and the process is represented as,

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \xrightarrow{A_{11}} B. \quad (\text{A4.1.15})$$

The above discussion shows that $M_{2|1}$ and $C_{22|1}$ can be obtained through the subsweep operation,

$$\begin{bmatrix} C_{11} & C_{12} & -E_{1.} \\ C_{21} & C_{22} & M_{2.} \end{bmatrix} \xrightarrow{C_{11}} [C_{22|1.} \quad M_{2|1.}], \quad (\text{A4.1.16})$$

regardless of whether C_{11} is positive definite or positive semidefinite.

Conditional Matrix-normal Inverted-Wishart and Matrix-T Distributions.

Proceeding in an obvious analogous way, the pivoting scheme for the matrix-normal can be expanded in order to obtain the parameters of the conditional Matrix-normal Inverted-Wishart in (A3.2.21) by augmenting the sweep operation (A4.1.13) into,

$$\begin{bmatrix} C_{11} & C_{12} & -E_{1.} \\ C_{21} & C_{22} & M_{2.} \\ E_{2.} & -M'_{2.} & S \end{bmatrix} \rightarrow \begin{bmatrix} C_{11}^{-1} & C_{11}^{-1}C_{12} & -C_{11}^{-1}E_{1.} \\ -C_{21}C_{11}^{-1} & C_{22|1.} & M_{2|1.} \\ -E_{2.}C_{11}^{-1} & -M'_{2|1.} & S_{|1.} \end{bmatrix}. \quad (\text{A4.1.17})$$

Of course, the general and more efficient subsweep operation (A4.1.16) can be expanded to,

$$\begin{bmatrix} C_{11} & C_{12} & -E_{1.} \\ C_{21} & C_{22} & M_{2.} \\ E_{2.} & -M'_{2.} & S \end{bmatrix} \xrightarrow{C_{11}} [C_{22|1.} \quad M_{2|1.}] \quad (\text{A4.1.18})$$

Furthermore, (A4.1.17) and (A4.1.18) also yield the Matrix-T conditional parameters appearing in (A3.2.23). Notice that a similar result for subsets of rows follows from (A3.2.26).

Care is necessary for updating the shape parameter d when applying (A4.1.18) since it must remain the same when a null pivot is skipped and must be increased by one unit when a pivotal operation (over a non-zero pivot) is performed, i.e. its final value is the initial plus the rank of C_{11} .

CHAPTER 5

PLUG-IN ESTIMATION, INFORMATION AND DYNAMIC LINEAR MODELS

A very convenient feature of Bayesian forecasting models is their parametric representation. This formulation provides a natural environment in which the modeller can express his/her beliefs about the time series in question. The Bayesian paradigm offers, in the form of the posterior density, the necessary feedback for the parameters of the model. For easy interpretation it is useful to summarize the parametric density by point estimates. A common choice is to use estimators based on standard loss functions such as quadratic, absolute and zero-one error loss. Very often estimates of functions of the parameters are of interest; for example, if the trend of a dynamic model contains harmonics we would like to have estimates for the amplitude and phases, and in the case of a multivariate time series the correlation and even the eigen-structure of the scale matrix are of concern. The usual Bayesian procedure consists in finding the distribution of the transformed parameters which provides the basis for the standard point estimation. Unfortunately, the resulting distributions are often difficult to work with analytically, as in the above examples.

This chapter is concerned with a plug-in estimation method and a related measure of information between observations and parameters in dynamic linear models. The loss function on which the plug-in estimates (PIE) are based is the Kullback and Liebler (1951) directed divergence of the estimated likelihood (using a plug-in rule, Dawid (1984)) from the actual likelihood, and reflects our interest in forecasting. A bonus of these estimates is that they are invariant under parametric transformations. In Section 5.1 the information, and plug-in estimators for DLM's are obtained. In Section 5.2 a dynamic multivariate regression is employed for modelling the energy consumption by primary fuel inputs and a use of PIE's is illustrated by estimating correlations across series. The decision theoretic support for the plug-in estimation method, together with the related measure of information between random variables, is presented in Appendix A5.1. Useful results about the entropy and information of some matrix-variate random variables are given in Appendix A5.2.

5.1 DYNAMIC LINEAR MODELS.

Plug-in estimators have two main uses in Bayesian forecasting. Firstly, they can be employed as convenient point estimators for (functions of) the parameters. Secondly, the estimated likelihood can be considered as a simple approximation of the predictive density. Henceforth we make use of the results and terminology of plug-in estimation and therefore it is assumed that the reader is familiar with them. These new results are derived in Appendix A5.1.

As an initial illustration we consider a very simple and yet the most widely used DLM: the steady (also called first order polynomial) model. The defining equations and distributions are,

$$y_t = \theta_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2), \quad (5.1a)$$

$$\theta_t = \theta_{t-1} + f_t, \quad f_t \sim N(0, \omega \sigma^2), \quad (5.1b)$$

$$\theta_{t-1} \sim N(m_{t-1}, c_{t-1} \sigma^2), \quad \sigma^2 \sim \Gamma^{-1}\left(\frac{1}{2}d_{t-1}, \frac{1}{2}s_{t-1}\right). \quad (5.1c)$$

The updating recurrences can be written as

$$c_t^* = \omega + c_{t-1}, \quad m_t^* = m_{t-1}, \quad (5.2a)$$

$$\hat{y}_t = 1 + c_t^*, \quad \hat{y}_t = m_t^*, \quad (5.2b)$$

$$c_t = \frac{c_t^*}{(1 + c_t^*)}, \quad m_t = \frac{1}{1 + c_t^*} m_t^* + \frac{c_t^*}{1 + c_t^*} \hat{y}_t, \\ s_t = s_{t-1} + \frac{(y_t - \hat{y}_t)^2}{1 + c_t^*}, \quad d_t = d_{t-1} + 1. \quad (5.2c)$$

Thus, in accordance with the results given in Appendix A5.1, the PIE's at time t (before observing y_t) are given by,

$$\hat{\theta}_t = m_t^* \quad \text{and} \quad \hat{\sigma}^2 = (1 + c_t^*) \frac{s_{t-1}}{d_{t-1} - 2}. \quad (5.3)$$

The first formula in (5.3) gives us yet another justification for the usual point estimator of θ_t . However, the estimator for σ^2 takes into account the uncertainty about θ_t and suggests that the usual plug-in recipe

$$\tilde{\sigma}^2 = \underset{\sigma^2}{\text{E}} \sigma^2 = \frac{s_{t-1}}{d_{t-1} - 2}$$

underestimates the variance when θ_t is unknown.

The loss function associated with the PIE's (5.3) is

$$L\left(\begin{bmatrix} \theta_t \\ \sigma^2 \end{bmatrix}, \begin{bmatrix} \hat{\theta}_t \\ \hat{\sigma}^2 \end{bmatrix}\right) = \frac{1}{2} \left(\frac{(\theta_t - \hat{\theta}_t)^2}{\hat{\sigma}^2} + \left(\frac{\sigma^2}{\hat{\sigma}^2} - 1 \right) + \log \frac{\hat{\sigma}^2}{\sigma^2} \right). \quad (5.4)$$

Notice that the square-error penalty in the estimation of θ is relative to the estimate of σ^2 , and this is precisely the reason why, in general, $\hat{\sigma}^2$ is different from the usual estimator. The EVPI, as defined in Section A5.1.1, is

$$\text{EVPI} = \frac{1}{2} (\log(1 + c_t^*) + \delta(\frac{1}{2}d_{t-1}) - \log(\frac{1}{2}d_{t-1} - 1)), \quad (5.5)$$

where δ denotes the digamma function; see Section 1.5.

The information between y_t and $\begin{bmatrix} \theta_t \\ \sigma^2 \end{bmatrix}$ is

$$I\left(y_t, \begin{bmatrix} \theta_t \\ \sigma^2 \end{bmatrix}\right) = \frac{1}{2} \log(1 + c_t^*) - \frac{1}{2} + \log \left(\frac{(\frac{1}{2}d_{t-1} - 1)!}{(\frac{1}{2}d_{t-1} - \frac{1}{2})!} \right) \\ + \left(\frac{1}{2}d_{t-1} + \frac{1}{2} \right) \delta\left(\frac{1}{2}d_{t-1} + \frac{1}{2}\right) - \frac{1}{2} d_{t-1} \delta\left(\frac{1}{2}d_{t-1}\right). \quad (5.6)$$

Hence the ENLE, as defined in Section A5.1.2, only depends upon the degrees of freedom. The EVPI, $I\left(y_t, \begin{bmatrix} \theta_t \\ \sigma^2 \end{bmatrix}\right)$ and ENLE, are computed in Table 5.1 for the model (5.1) with $c_0 = \epsilon^{-1}$, $d_0 = \epsilon = 10^{-5}$ and $w = 1$. The non-zero limiting behaviour of the EVPI and the $I\left(y_t, \begin{bmatrix} \theta_t \\ \sigma^2 \end{bmatrix}\right)$ is typical of a dynamic model, the reason is that the limiting value of c_t^* , the reciprocal of the golden section $\left(\frac{\sqrt{5}+1}{2}\right)$ in this particular example, is non-zero. On the other hand, the ENLE becomes rapidly negligible relative to the EVPI, and reflects the fact that the normal distribution approximates very well the t distribution as the degrees of freedom increase. In fact it is shown in Subsection 5.1.1 that the last column in Table 5.1 is still valid for any univariate DLM. Notice that non-zero asymptotic information means that, in a genuine dynamic model, we can not afford to stop looking at new data.

t	c^*	d	EVPI	INFO	ENLE
4	1.6250	3	0.8474	0.6526	0.194767
5	1.6190	4	0.6928	0.6090	0.083752
6	1.6182	5	0.6301	0.5832	0.046849
7	1.6181	6	0.5960	0.5661	0.029950
8	1.6180	7	0.5746	0.5538	0.020796
9	1.6180	8	0.5600	0.5447	0.015281
10	1.6180	9	0.5493	0.5376	0.011702
11	1.6180	10	0.5411	0.5319	0.009248
12	1.6180	11	0.5347	0.5272	0.007492
13	1.6180	12	0.5296	0.5234	0.006193
14	1.6180	13	0.5253	0.5201	0.005204
15	1.6180	14	0.5217	0.5173	0.004435
16	1.6180	15	0.5187	0.5149	0.003824
17	1.6180	16	0.5161	0.5127	0.003331
18	1.6180	17	0.5138	0.5109	0.002928
19	1.6180	18	0.5118	0.5092	0.002594
20	1.6180	19	0.5100	0.5077	0.002314
21	1.6180	20	0.5085	0.5064	0.002077
22	1.6180	21	0.5071	0.5052	0.001874
23	1.6180	22	0.5058	0.5041	0.001700
24	1.6180	23	0.5046	0.5031	0.001549
25	1.6180	24	0.5036	0.5022	0.001417

Table 5.1 EVPI, Information and ENLE for the Steady Model (5.1).

5.1.1 Dynamic Weighted Multivariate Regression.

The PIE's for the DWMR model (3.32) can be easily derived by noticing that the solution (3.18) to the problem (3.16) implies in particular that the Kullback-Liebler directed divergence of the distribution of a random vector from a multivariate normal is minimized by setting the first and second moments

of the normal approximation to be the same as those of the random vector (the standard recipe!). Therefore, from (3.32a) the PIE's must satisfy,

$$\underline{x}' \hat{\Theta}_t = \underline{x}' M_t^* \quad \text{and} \quad v_t \hat{\Sigma} = (v_t + \underline{x}' C_t^* \underline{x}_t) \frac{S_{t-1}}{d_{t-1} - 2}, \quad (5.7)$$

where $M_t^* = G_t M_{t-1}$ and $C_t^* = W_t + G_t C_{t-1} G_t'$. In particular, the solution,

$$\hat{\Theta}_t = M_t^* \quad \text{and} \quad \hat{\Sigma} = (1 + \frac{1}{v_t} \underline{x}' C_t^* \underline{x}_t) \frac{S_{t-1}}{d_{t-1} - 2}, \quad (5.8)$$

is optimal for any \underline{x}_t . Notice again that the estimator of Σ is the mean value $\frac{S_{t-1}}{d_{t-1} - 2}$ modified by a correction factor $(1 + \frac{1}{v_t} \underline{x}' C_t^* \underline{x}_t)$. This correction factor accounts for the uncertainty about Θ_t .

Of course, updated and/or predictive PIE's can be obtained from their corresponding distributions. For instance, after observing \underline{y}_t the PIE's for the trend $\underline{x}' \Theta_t$ and the variance $v_t \Sigma$ are,

$$\underline{x}' \hat{\Theta}_t = \underline{x}' M_t \quad \text{and} \quad v_t \hat{\Sigma} = (v_t + \underline{x}' C_t \underline{x}_t) \frac{S_t}{d_t - 2}. \quad (5.9)$$

In addition, the PIE for the long-term trend $\underline{x}'_{t+s} \theta_{t+s}$ is merely the forecasting function,

$$\underline{x}'_{t+s} \hat{\Theta}_{t+s} = \underline{x}'_{t+s} M_{t+s} = F_t(s), \quad (5.10)$$

where M_{t+s} is defined recursively as $M_{t+s} = G_{t+s} M_{t+s-1}$, for $s = 1, \dots$

In general, the optimal solution for $\hat{\Theta}_t$ is not unique. However, $\hat{\Theta}_t = M_t$ is the only one valid for any \underline{x}_t . Furthermore, for a noise-free constant DWMR, $\hat{\Theta} = M_t$ is the unique solution if and only if the DWMR is observable, since (5.10) implies, $\underline{x}' G^s \hat{\Theta}_t = \underline{x}' G^s M_t$ for $s = 0, \dots, p-1$, i.e. $T_{\underline{h}} \hat{\Theta}_t = T_{\underline{h}} M_t$ where

$$T_{\underline{h}} = \underline{h} \begin{bmatrix} \underline{x}' \\ \underline{x}' G \\ \vdots \\ \underline{x}' G^{p-1} \end{bmatrix}$$

is a full rank matrix for some \underline{h} .

The loss function associated with the PIE's (5.8) is

$$l\left(\begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix}, \begin{bmatrix} \hat{\Theta}_t \\ \hat{\Sigma} \end{bmatrix}\right) = \frac{1}{2} \left(\frac{1}{v_t} \underline{x}' \underline{x}_t \text{tr}((\Theta_t - \hat{\Theta}_t)' (\Theta_t - \hat{\Theta}_t) \hat{\Sigma}^{-1}) + \text{tr}(\Sigma \hat{\Sigma}^{-1}) - q + \log \frac{|\hat{\Sigma}|}{|\Sigma|} \right). \quad (5.11)$$

The EVPI is given by,

$$\text{EVPI} = \frac{1}{2} (q \log(1 + \frac{1}{v_t} \underline{x}' C_t^* \underline{x}_t) + \sum_{j=1}^q \delta((\frac{1}{2}(d_{t-1} + q - j)) - q \log(\frac{1}{2} d_{t-1} - 1))). \quad (5.12)$$

The information is,

$$\begin{aligned} I\left(\underline{y}_t, \begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix}\right) &= \frac{1}{2} q \log(1 + \frac{1}{v_t} \underline{x}' C_t^* \underline{x}_t) - \frac{1}{2} q + \sum_{j=1}^q \log \left(\frac{(\frac{1}{2}(\nu_{t-1} - j - 1))!}{(\frac{1}{2}(\nu_{t-1} - j))!} \right) \\ &\quad + \frac{1}{2} (\nu_{t-1} + 1) \sum_{j=1}^q \delta(\frac{1}{2}(\nu_{t-1} - j + 2)) - \frac{1}{2} \nu_{t-1} \sum_{j=1}^q \delta(\frac{1}{2}(\nu_{t-1} - j + 1)) \end{aligned} \quad (5.13)$$

where $\nu_{t-1} = d_{t-1} + q - 1$. Formulas (5.11-5.13) are generalizations of (5.4-5.6) and particular cases of (5.18, 5.19 and 5.22). From (5.12) and (5.13) it is evident that the ENLE only depends upon the degrees of freedom. Furthermore, the ENLE for any univariate DLM is the same as that of model (5.1) and in that sense the last column of Table 5.1 is rather general.

From (5.12) it is not difficult to see that the EVPI limiting value for a constant model is,

$$\lim_{t \rightarrow \infty} \text{EVPI} = \frac{1}{2}q \log(1 + \frac{1}{v} \underline{x}' C^* \underline{x}) \quad \text{where } C^* = \lim_{t \rightarrow \infty} C_t^*. \quad (5.14)$$

Moreover, the $I\left(\underline{y}_t, \begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix}\right)$ limiting value is also given by the right-hand side of (5.14) since (A5.14) and (5.13) imply $\text{EVPI} \geq I\left(\underline{y}_t, \begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix}\right) \geq \frac{1}{2}q \log(1 + \frac{1}{v} \underline{x}' C_t^* \underline{x})$. Formula (5.14) is useful for looking at the asymptotic behaviour in terms of information for a particular choice of driving variances v and W (or a set of discount factors).

Notice that the above discussion justifies the use of an estimated (plug-in) likelihood for predictive purposes; as time passes the ENLE tends to zero.

5.1.2 Dynamic Linear Matrix-variate Regression.

The optimization problem for finding the PIE's associated with the DLMR model (3.5),

$$\max_{\hat{\Theta}_t, \hat{\Sigma}} E_{\Theta_t, \Sigma} U\left(\begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix}, \begin{bmatrix} \hat{\Theta}_t \\ \hat{\Sigma} \end{bmatrix}\right), \quad (5.15)$$

is a particular case of (3.16). To see this, we note that problem (5.15) can be restated as,

$$\max_{\hat{\Theta}_t, \hat{\Sigma}} E_{Y_t} \log p(Y_t | \hat{\Theta}_t, \hat{\Sigma}), \quad (5.16)$$

where the probability density $p(Y_t | \hat{\Theta}_t, \hat{\Sigma})$ corresponds to the estimated distribution $Y_t \sim N(X_t \hat{\Theta}_t, V_t, \hat{\Sigma})$, and the expected value is taken over the target predictive distribution $Y_t \sim T(\hat{Y}_t, \check{Y}_t, S_{t-1}, d_{t-1})$, in accordance with (3.10b). Therefore, from the general solution (3.18) and (A5.2.1b) we obtain,

$$\hat{\Theta}_t = M_t^* \quad \text{and} \quad \hat{\Sigma} = (1 + \frac{1}{r} \text{tr}(V_t^{-1} X_t C_t^* X_t')) \frac{S_{t-1}}{d_{t-1} - 2}. \quad (5.17)$$

These estimators are of course a generalization of (5.8). Comments similar to those following (5.8), regarding the uniqueness of the PIE's and updated/predictive PIE's apply with obvious modifications to (5.17).

The corresponding loss function is,

$$I\left(\begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix}, \begin{bmatrix} \hat{\Theta}_t \\ \hat{\Sigma} \end{bmatrix}\right) = \frac{1}{2} (\text{tr}((\Theta_t - \hat{\Theta}_t)' X_t' V_t^{-1} X_t (\Theta_t - \hat{\Theta}_t) \hat{\Sigma}^{-1})) + r(\text{tr}(\Sigma \hat{\Sigma}^{-1}) - q + \log(|\hat{\Sigma}| |\Sigma|^{-1})). \quad (5.18)$$

This generalization of (5.11) may be verified with the aid of formulas (A5.1.2), (A3.2.10) and (A5.2.1a). Furthermore, the EVPI is given by,

$$\text{EVPI} = \frac{1}{2}r(q \log(1 + \frac{1}{r} \text{tr}(V_t^{-1} X_t C_t^* X_t') + \sum_{j=1}^q \delta(\frac{1}{2}(d_{t-1} + q - j)) - q \log(\frac{1}{2}d_{t-1} - 1)). \quad (5.19)$$

This formula can be shown as follows. From (5.17) and (A5.2.1b) it is not difficult to see that,

$$\mathbb{E}_{\Theta_t, \Sigma} \text{tr}((\Sigma + \frac{1}{r}(\Theta_t - \hat{\Theta}_t)' X_t' V_t^{-1} X_t (\Theta_t - \hat{\Theta}_t)) \hat{\Sigma}^{-1}) = q. \quad (5.20)$$

Therefore, from (5.18)

$$\mathbb{E}_{\Theta_t, \Sigma} I\left(\begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix}, \begin{bmatrix} \hat{\Theta}_t \\ \hat{\Sigma} \end{bmatrix}\right) = \frac{1}{2}r(\log |\hat{\Sigma}| - \mathbb{E} \log |\Sigma|), \quad (5.21)$$

which, according to (5.17) and (A5.2.2), results in (5.19).

The information between Y_t and $\begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix}$ is given by,

$$\begin{aligned} I(Y_t, \begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix}) &= \frac{1}{2}q(\log |I + V_t^{-1} X_t C_t^* X_t'|) - \frac{1}{2}rq + \sum_{j=1}^q \log \left(\frac{(\frac{1}{2}(\nu_{t-1} - j - 1))!}{(\frac{1}{2}(\nu_{t-1} + r - j - 1))!} \right) \\ &\quad + \frac{1}{2}(\nu_{t-1} + r) \sum_{j=1}^q \delta(\frac{1}{2}(\nu_{t-1} + r - j + 1)) - \frac{1}{2}\nu_{t-1} \sum_{j=1}^q \delta(\frac{1}{2}(\nu_{t-1} - j + 1)), \end{aligned} \quad (5.22)$$

where $\nu_{t-1} = d_{t-1} + q - 1$. Equation (5.22) may be verified as follows. From (A5.2.10) we have,

$$\begin{aligned} I(Y_t, \begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix}) &= -\frac{1}{2}qr + \sum_{j=1}^q \log \left(\frac{(\frac{1}{2}(\nu_{t-1} - j - 1))!}{(\frac{1}{2}(\nu_{t-1} + r - j - 1))!} \right) \\ &\quad + \frac{1}{2}(\nu_{t-1} + r) \sum_{j=1}^q \delta(\frac{1}{2}(\nu_{t-1} + r - j + 1)) - \frac{1}{2}\nu_{t-1} \sum_{j=1}^q \delta(\frac{1}{2}(\nu_{t-1} - j + 1)), \end{aligned} \quad (5.23)$$

Now, $I(Y_t, \begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix})$ can be written as,

$$I(Y_t, \begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix}) = \mathbb{E}_{\Sigma} \left[\mathbb{E}_{\begin{bmatrix} \Theta_t \\ \Sigma \end{bmatrix} | \Sigma} \log \left(\frac{p(\Theta_t | Y_t, \Sigma)}{p(\Theta_t | \Sigma)} \right) \right] + I(Y_t, \Sigma). \quad (5.24)$$

where, according to (A5.2.6) and (4.4), the first term in the right-hand side of (5.24) is given by,

$$\frac{1}{2}q \log \frac{|C_t^*|}{|C_t|} = \frac{1}{2}q \log |I + C_t^* X_t' V_t^{-1} X_t| = \frac{1}{2}q \log |I + V_t^{-1} X_t C_t^* X_t'|. \quad (5.25)$$

Thus, formula (5.22) follows immediately from (5.23-25). Moreover, following a similar pattern, it is not difficult to show that the counterpart of formula (5.23) is given by,

$$\begin{aligned} I(Y_t, \Theta_t) &= \frac{1}{2}q \log |I + V_t^{-1} X_t C_t^* X_t'| + \sum_{j=1}^q \log \left(\frac{(\frac{1}{2}(\nu_{t-1} + r + p - j - 1))! ((\frac{1}{2}(\nu_{t-1} - j - 1))!)^2}{(\frac{1}{2}(\nu_{t-1} + p - j - 1))! ((\frac{1}{2}(\nu_{t-1} + r - j - 1))!)^2} \right) \\ &\quad - \frac{1}{2}(\nu_{t-1} + r + p) \sum_{j=1}^q \delta(\frac{1}{2}(\nu_{t-1} + r + p - j + 1)) + \frac{1}{2}(\nu_{t-1} + p) \sum_{j=1}^q \delta(\frac{1}{2}(\nu_{t-1} + p - j + 1)) \\ &\quad + (\nu_{t-1} + r) \sum_{j=1}^q \delta(\frac{1}{2}(\nu_{t-1} + r - j + 1)) - \nu_{t-1} \sum_{j=1}^q \delta(\frac{1}{2}(\nu_{t-1} - j + 1)). \end{aligned}$$

It is interesting to note that as opposed to the DWMR case, in general, the limiting value of ENLE is non-zero. Let us consider the multivariate extension of the DLM (2.15), i.e. a DLMR with $q = 1$,

$$\underline{y}_t = X_t \underline{\theta}_t + \underline{e}_t, \quad \underline{e}_t \sim N(\underline{0}, \sigma^2 V_t), \quad (5.26a)$$

$$\underline{\theta}_t = G_t \underline{\theta}_{t-1} + \underline{f}_t, \quad \underline{f}_t \sim N(\underline{0}, \sigma^2 W_t), \quad (5.26b)$$

$$\underline{\theta}_{t-1} \sim N(\underline{m}_{t-1}, \sigma^2 C_{t-1}), \quad \sigma^2 \sim \Gamma^{-1}(\frac{1}{2}d_{t-1}, \frac{1}{2}s_{t-1}). \quad (5.26c)$$

Then the EVPI and $I\left(\underline{y}_t, \left[\begin{array}{c} \underline{\theta}_t \\ \sigma^2 \end{array}\right]\right)$ are given by,

$$EVPI = \frac{1}{2}r(\log(1 + \frac{1}{r} \text{tr}(V_t^{-1} X_t C_t^* X_t') + \delta(\frac{1}{2}d_{t-1}) - \log(\frac{1}{2}d_{t-1} - 1)), \quad (5.27)$$

$$I\left(\underline{y}_t, \left[\begin{array}{c} \underline{\theta}_t \\ \sigma^2 \end{array}\right]\right) = \frac{1}{2} \log |I + V_t^{-1} X_t C_t^* X_t'| - \frac{1}{2}r + \log \left(\frac{(\frac{1}{2}d_{t-1} - 1)!}{(\frac{1}{2}(d_{t-1} + r) - 1)!} \right) + \frac{1}{2}(d_{t-1} + r)\delta(\frac{1}{2}(d_{t-1} + r)) - \frac{1}{2}d_{t-1}\delta(\frac{1}{2}d_{t-1}). \quad (5.28)$$

Hence, the ENLE limiting value for a constant model is,

$$\frac{1}{2} \log \frac{(1 + \frac{1}{r} \text{tr}(V^{-1} X C^* X'))^r}{|I + V^{-1} X C^* X'|} \geq 0. \quad (5.29)$$

Moreover, the equality holds if and only if $V^{-1} X C^* X' = \lambda I$ for some λ . This means that, apart from special cases, the estimated likelihood is incapable of approximating the predictive density, and therefore the plug-in estimators must not be employed for these purposes in such circumstances.

5.2 EXAMPLE: ENERGY CONSUMPTION BY PRIMARY FUEL INPUTS.

We present here an application of the DWMR for modelling the UK Inland energy consumption by primary fuel inputs as in Quintana (1985). Then the use of PIE's as point estimates is illustrated by estimating the correlations across several time series.

5.2.1 Dataset.

The time series consists of 64 observations of four series each containing the monthly UK Inland energy consumption by primary fuel inputs (coal, petrol, gas and nuclear) measured in millions of tonnes of coal or coal equivalent, from 1979 to September 1984 inclusive (CSO Monthly Digest). These series are plotted in Figure 5.1. The coal and petrol consumptions overlap until the effects of the Miners' Strike are visible, then coal consumption declines and petrol takes over. Also apparent is the seasonality of the series (presumably the effect of temperature). This is particularly noticeable for gas consumption.

5.2.2 The Model.

A DWMR is employed for modelling the natural logarithms of the original series. A fortunate consequence of this is that inferences about the proportions of energy consumption by primary fuel input can be made virtually at once due to the relationship between the multivariate logistic-normal and log-normal distributions; see Section 7.2.

The dynamic model is given by (3.31), with following setting,

$$\begin{aligned}\underline{x}'_t &= [1, 0, 1, 0, 1, 0], \\ v_t &= v = 2.5, \\ G_t &= G = \text{diag} \left(\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} \cos \omega_1 & \sin \omega_1 \\ -\sin \omega_1 & \cos \omega_1 \end{bmatrix}, \begin{bmatrix} \cos \omega_2 & \sin \omega_2 \\ -\sin \omega_2 & \cos \omega_2 \end{bmatrix} \right), \\ \text{where } \omega_1 &= \frac{2\pi}{12}, \quad \omega_2 = \frac{2\pi}{3}, \\ W_t &= W = 2.5 \underline{w} \underline{w}', \quad \underline{w}' = (.1, .1, .02, .02, .02, .02), \\ \text{for } t &= 1, 2, \dots, 64,\end{aligned}$$

and a vague prior with $\epsilon = 10^{-5}$; see Section 7.1.

This setting means that the model is, in fact, a constant dynamic (non-weighted) multivariate regression, i.e. the model (3.31) with constant \underline{x}, v, G and W . The model is build-up via the superposition principle; see Subsection 2.3.1. The first block of G_t represents a linear trend; see Subsection 2.3.2. The second and third blocks represent harmonic trend; see Subsection 2.3.3. Thus, according to the discussion in Section 3.7, the dependent variables have their own dynamic trend consisting of a superposition of a linear trend plus two harmonics, with yearly and quarterly periods. The seasonal trend parameters, though dynamic, are almost noiseless in comparison with the linear trend parameters. This conveys the general suggestion that a seasonal trend is relatively stable over time in comparison with the deseasonalized (linear in this example) trend. Moreover, the variance across series (given Θ and Σ) is 2.5Σ (the value 2.5 is completely arbitrary).

This model was kept very simple and no attempt at system intervention was made whatsoever (to avoid hindsight). In practice, however, a more realistic approach might be to choose V_t and W_t according to the information available, and to introduce new parameters in order to explain the effects of special events such as the Miners' Strike.

5.2.3 Performance.

The merits of the dynamic model can be assessed in comparison with its static counterpart: the same model but with $W = O$.

The crosses in Figures 5.2-5 show the one-step forecasts (the predictive means in log-scale) using the dynamic model. In contrast, Figure 5.6 shows the coal consumption forecasts produced by the static model and its non-dynamic nature is evident from the poor predictions for $t > 58$. The discrimination between these rival models can be made via the multi-process models class I of Subsection 3.4.1. The log-odds of these alternative models are shown in figure 5.7 (assuming even prior odds); clearly the dynamic model outperforms the static one long before the Miners' Strike.

A similar procedure, described in Section 3.6, can be used for testing the null hypothesis H_0 : coal consumption is independent of the rest (given the parameters). Figure 5.8 displays the full/null model log-odds assuming even prior odds as before. Although the weight of evidence against H_0 is quite visible,

there is a strange behaviour at the end of the period; to understand it the PIE's of the correlations give us some insight.

5.2.4 On-line Estimates of the Correlations.

The updated PIE of Σ at time t is given by,

$$\hat{\Sigma} = \left(1 + \frac{1}{v} \underline{x}' C_t \underline{x}\right) \frac{S_t}{d_t - 2}, \quad (5.30)$$

in accordance with the comment that follows formula (5.8). Therefore, in view of the invariance property, in order to estimate the correlation matrix R associated with Σ , the value of $\hat{\Sigma}$ may be substituted into the defining equations,

$$R = \Delta^{-1} \Sigma \Delta^{-1} \quad \text{and} \quad \Delta = \text{diag}(\sigma_{11}^{\frac{1}{2}}, \dots, \sigma_{qq}^{\frac{1}{2}}), \quad (5.31)$$

i.e.

$$\hat{R} = [\hat{\rho}_{ij}] \quad \text{where} \quad \hat{\rho}_{ij} = \frac{S_{ijt}}{(S_{iit} S_{jst})^{\frac{1}{2}}}, \quad \text{for } i, j = 1, \dots, q. \quad (5.32)$$

Note that the estimates of the correlations do not depend on the correction factor $(1 + \frac{1}{v} \underline{x}' C \underline{x})$. Furthermore, $\hat{\rho}_{ij}$ are also the plug-in estimates of the correlations across series, i.e. the correlation associated with 2.5Σ . Of course, they can be interpreted directly as the predictive correlations.

The on-line estimates of the correlations between coal consumption and the rest are plotted in Figure 5.9. A quick look suggests that the homoscedasticity implicit in the model is a sensible assumption for $t \leq 58$, then the correlations change abruptly. This can be interpreted as a consequence of substituting coal consumption by the other fuels, particularly by petrol consumption. A means, already mentioned, of overcoming this structural change in Σ , is to expand the set of regressor parameters at time 58 and/or reflect our uncertainty about Σ by resetting the hyperparameters S_t and d_t to low values. This latter solution can be carried out systematically simulating a random walk type of evolution for Σ . The procedure is discussed in Subsection 6.3.2.

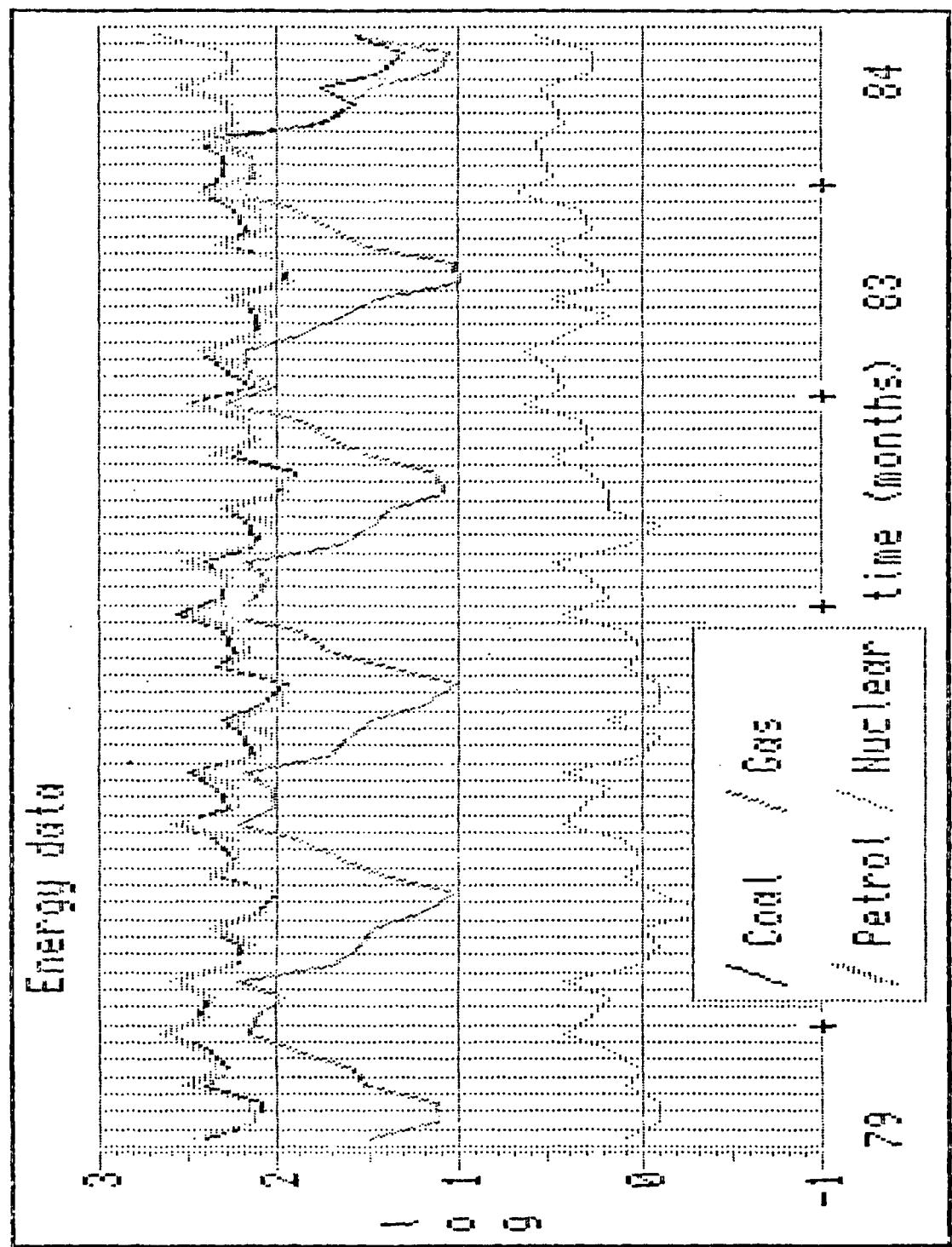


Figure 5.1 UK Inland energy consumption by primary fuel inputs: coal, petrol, gas and nuclear.

Coal data and forecasts

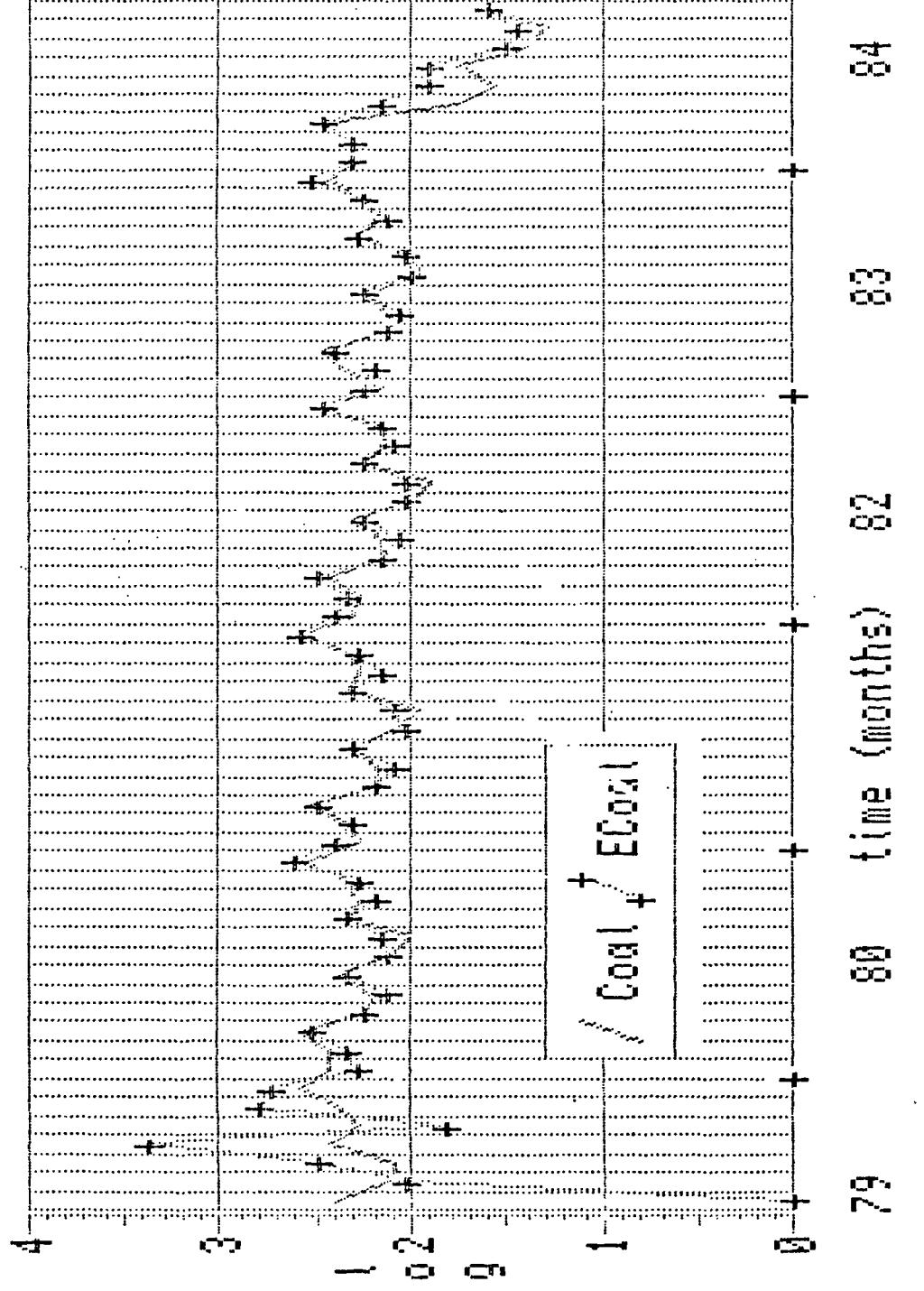


Figure 5.2 One-step forecasts for coal consumption.

Petrol data and forecasts

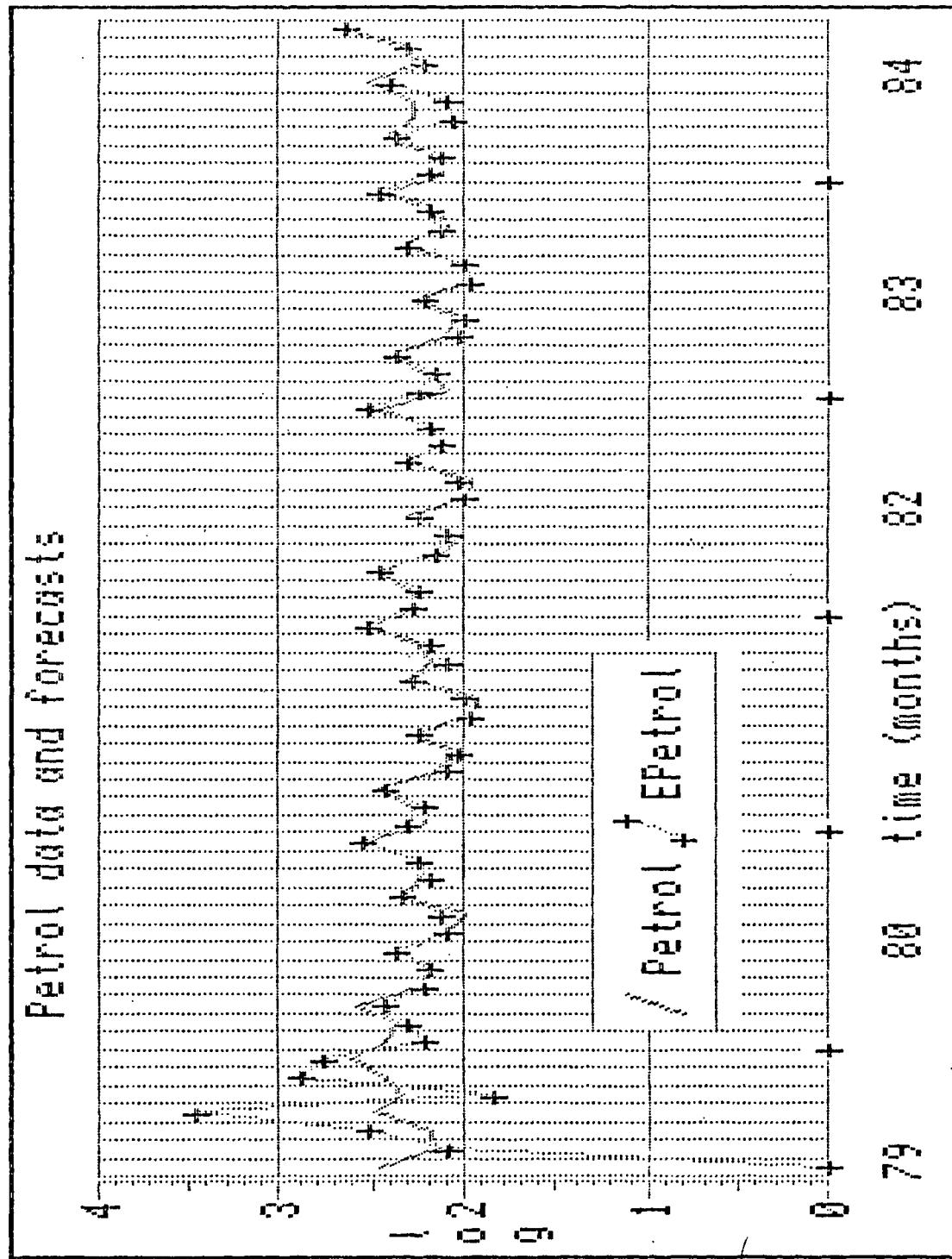


Figure 5.3 One-step forecasts for petrol consumption.

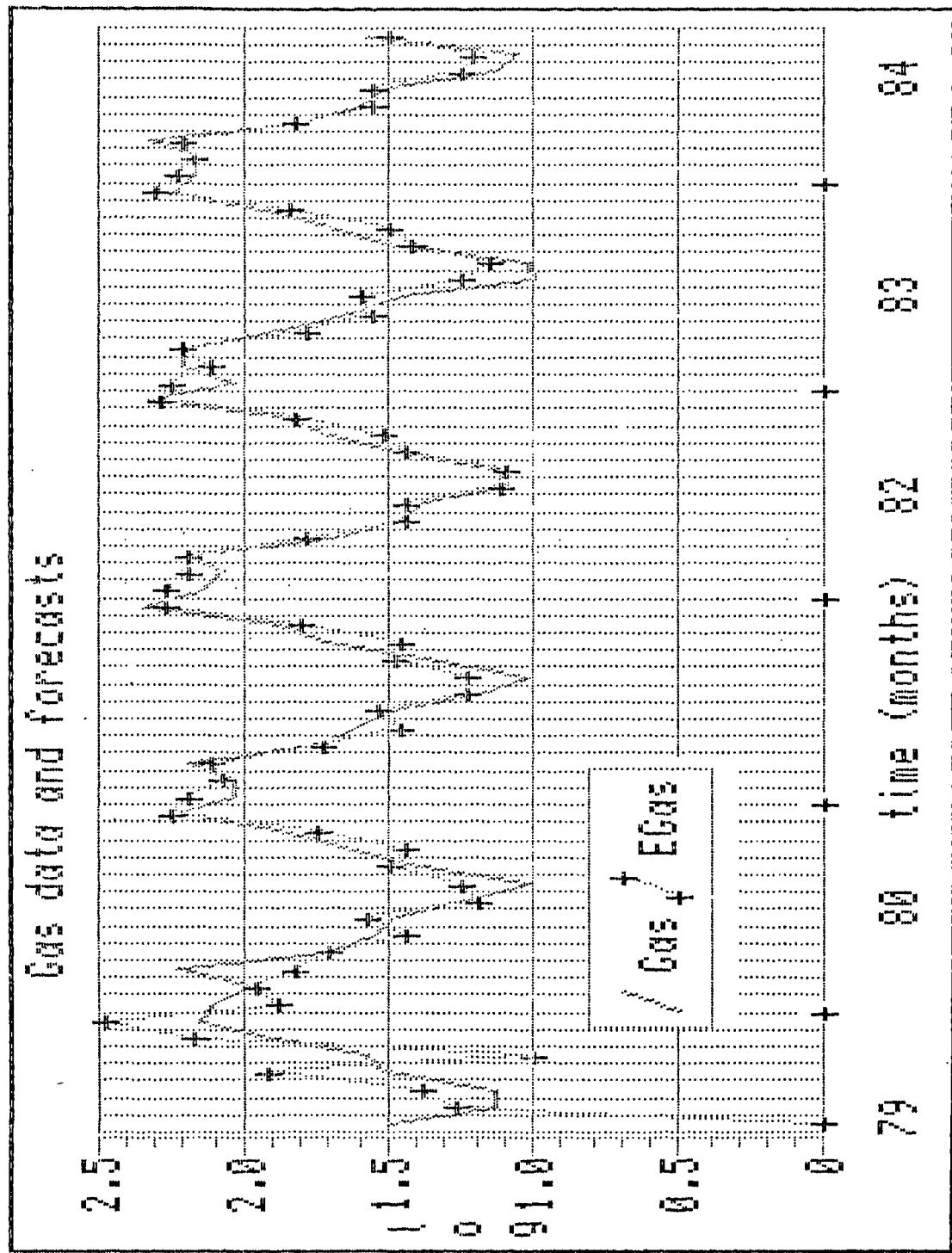


Figure 5.4 One-step forecasts for gas consumption.

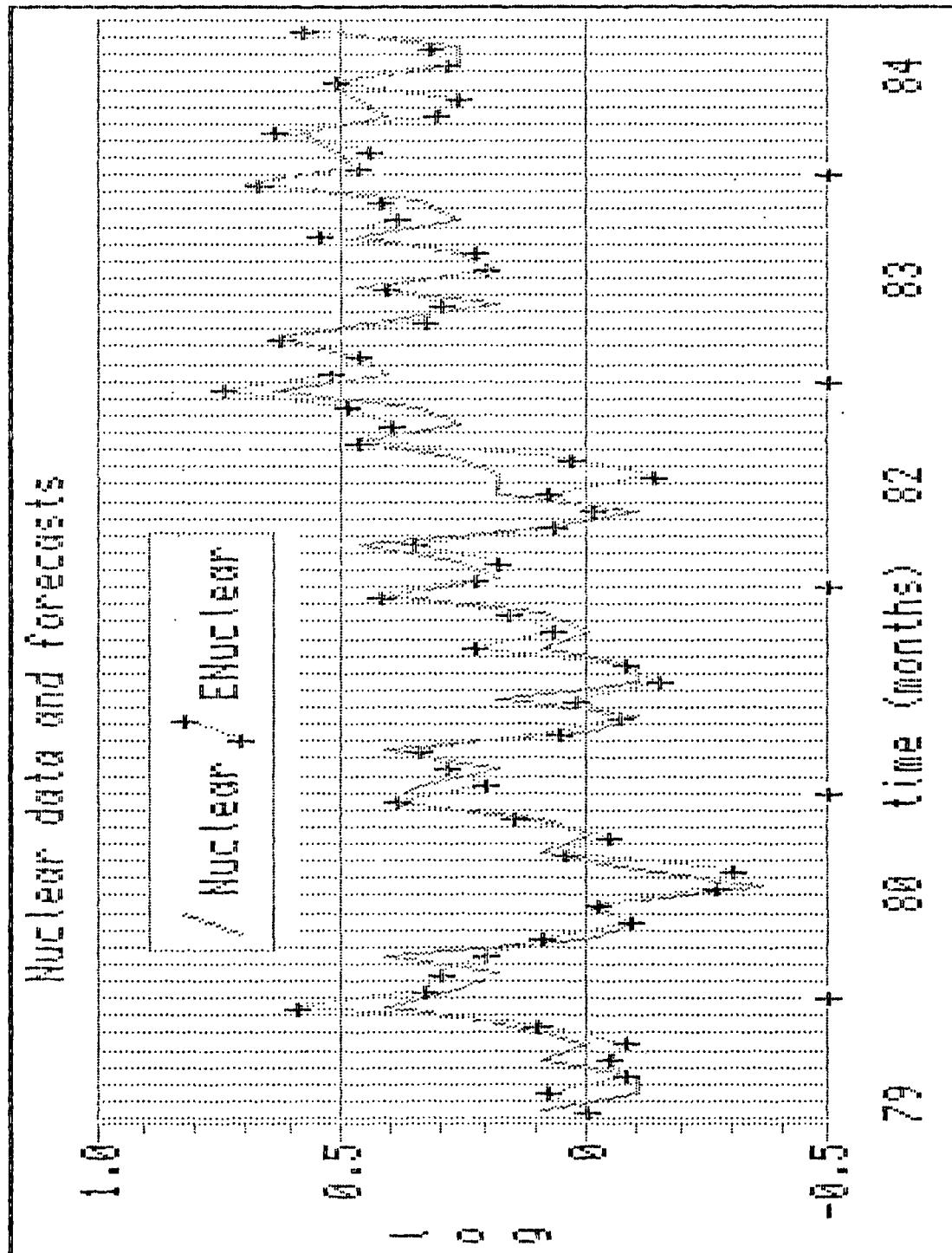


Figure 5.5 One-step forecasts for nuclear consumption.

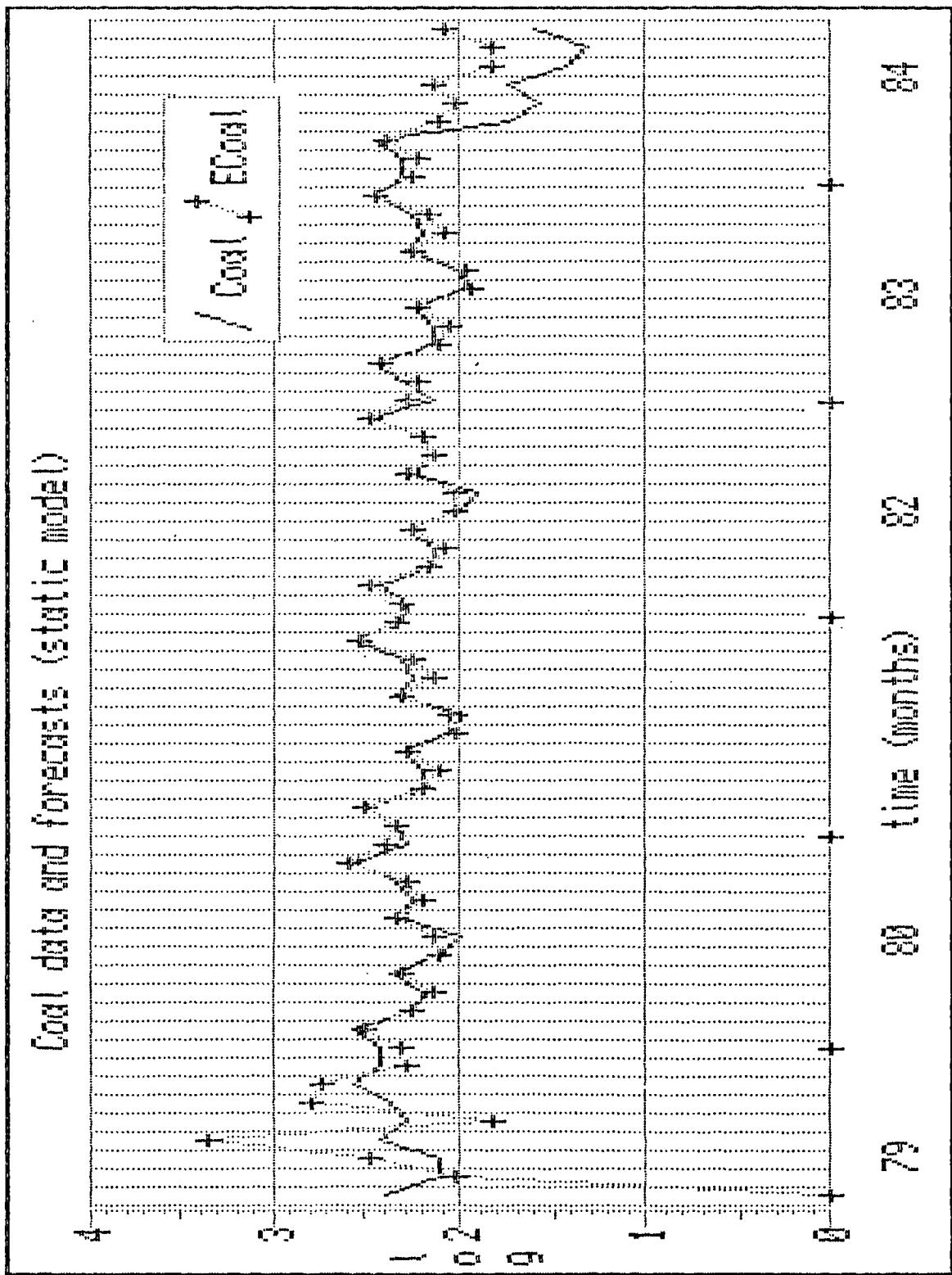


Figure 5.6 One-step forecasts for coal consumption (static model).

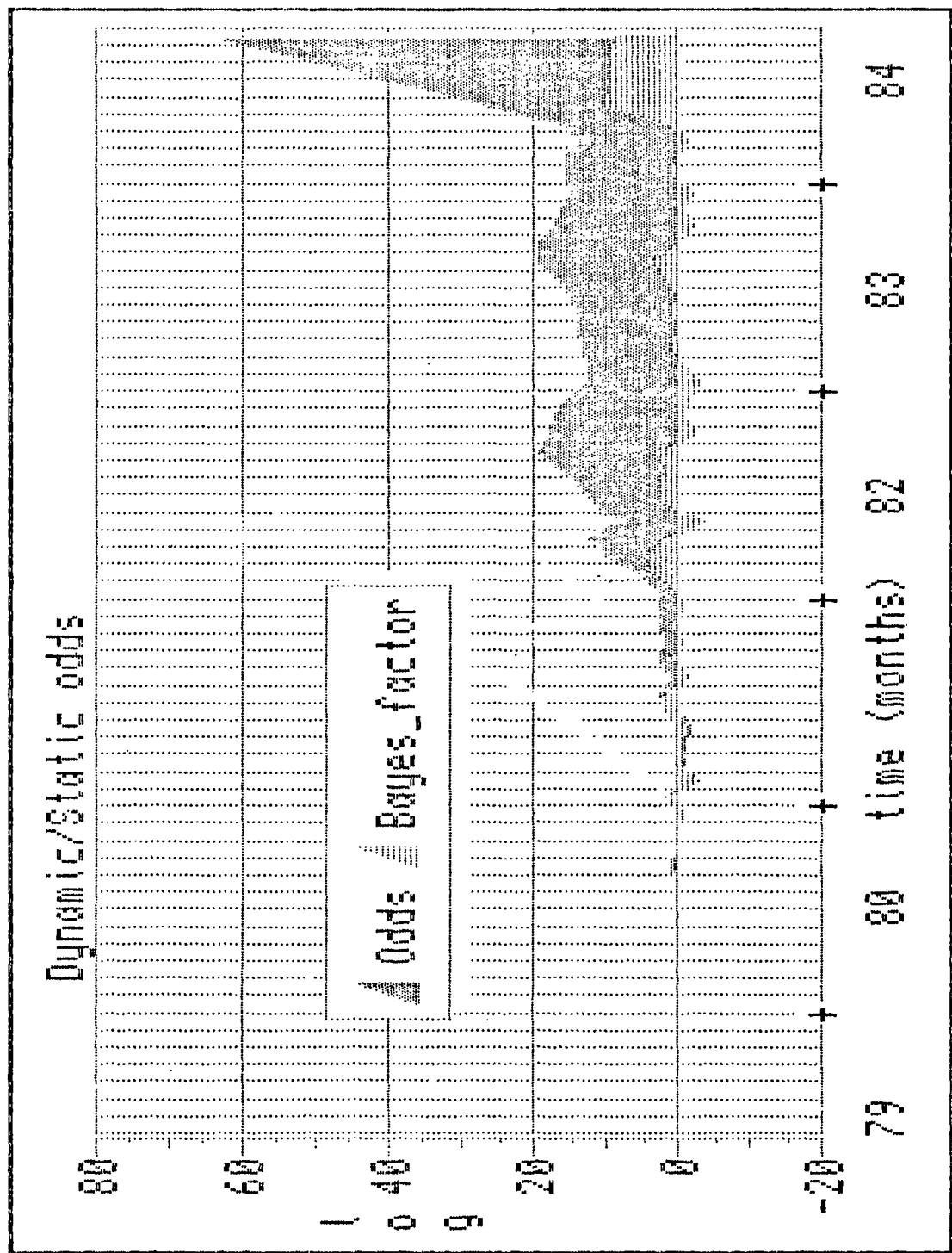


Figure 5.7 Dynamic model versus static model odds (in log-scale).

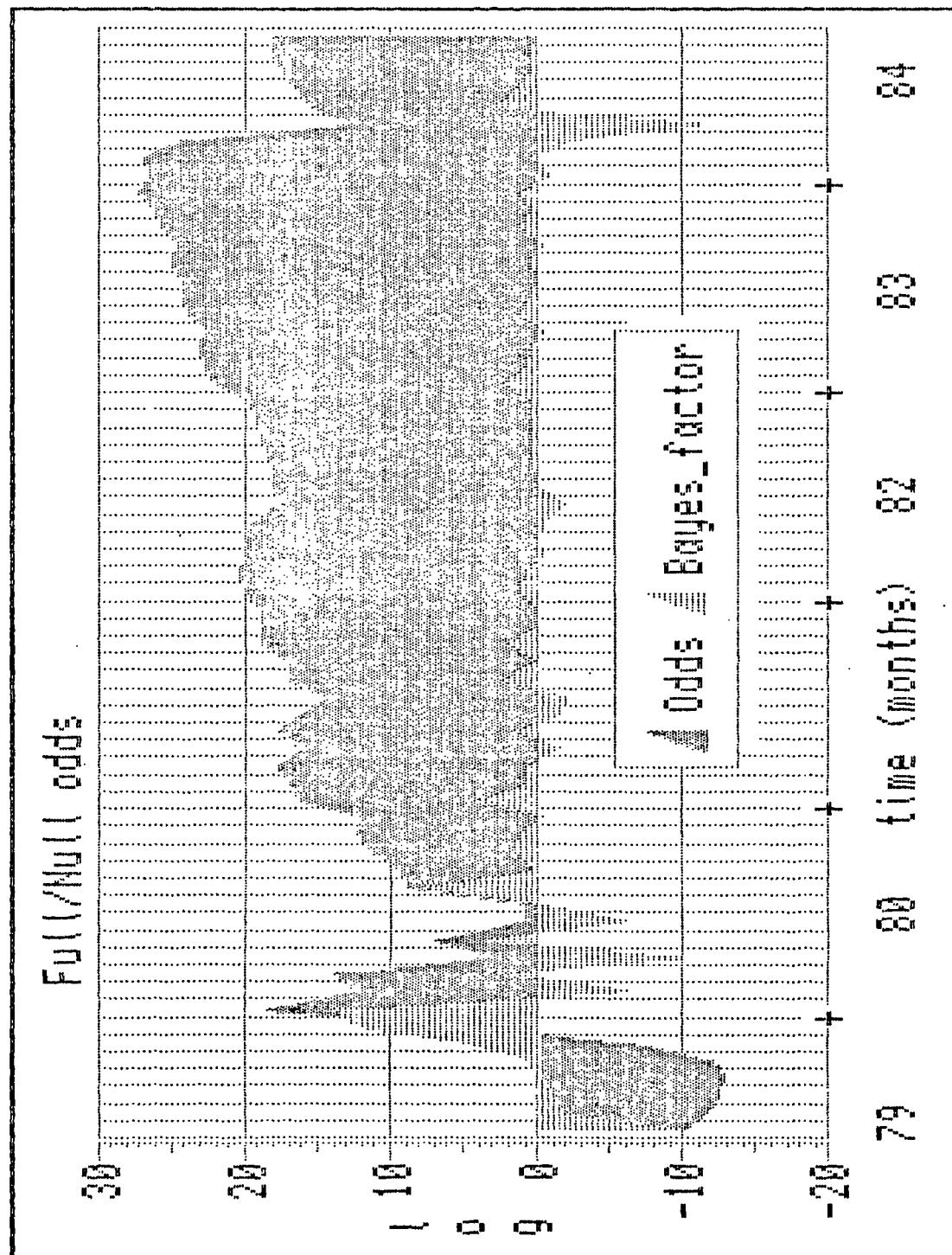


Figure 5.8 Full model versus null model odds (in log-scale).

Correlations between coal and the rest

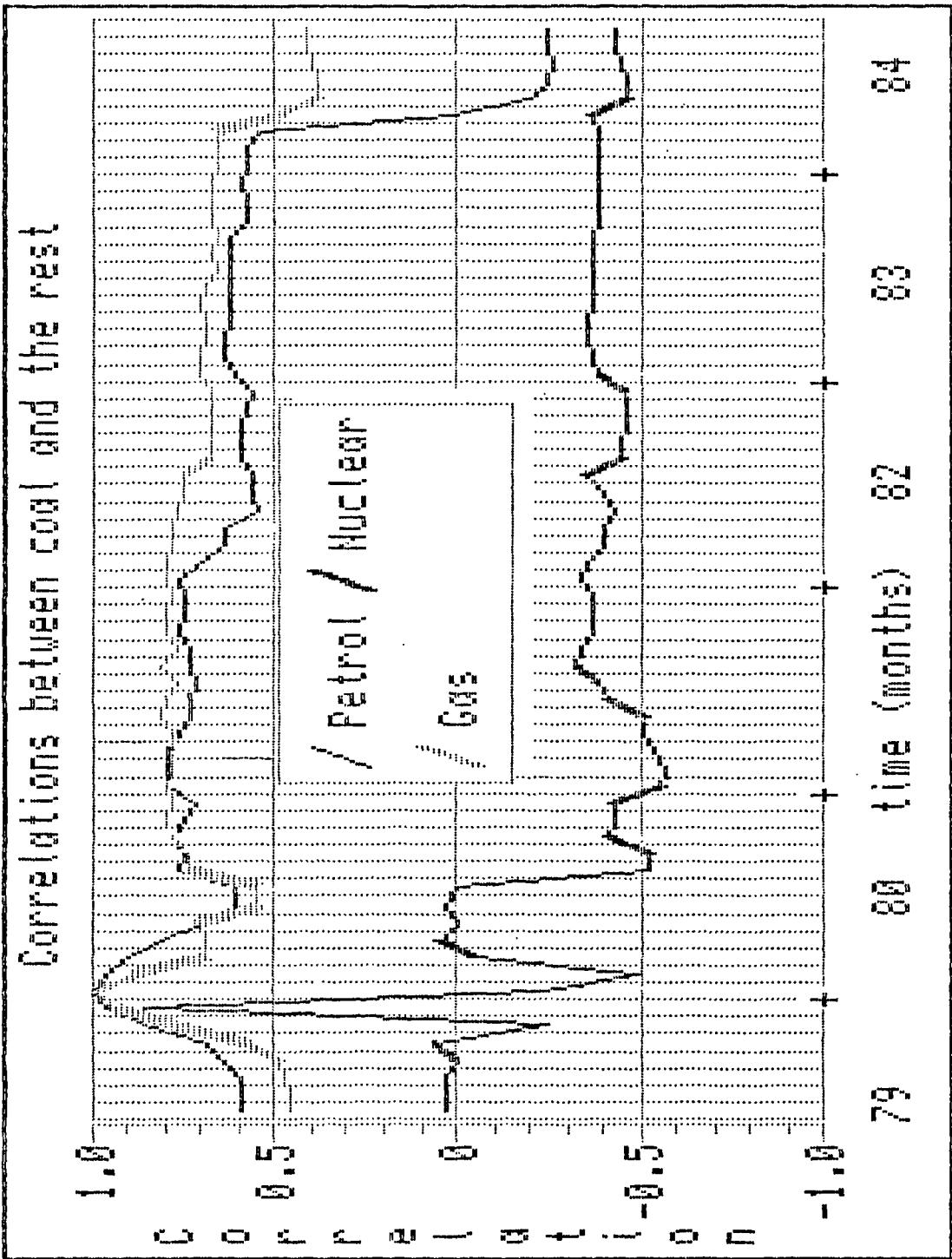


Figure 5.9 Plug-in estimates of the correlations between coal consumption and the rest.

APPENDIX A5.1.

PLUG-IN ESTIMATION AND INFORMATION.

A Bayesian modeller is forced to produce point estimators of the parameters for forecasting purposes. Here this problem is addressed using the Bayesian decision theoretic approach (e.g. Raiffa and Schlaifer, 1961; Berger, 1985), with the logarithmic scoring rule (LSR) as the utility function. The estimators obtained using this procedure resemble the maximum likelihood estimators in the sense that they are invariant under one-to-one transformations of the parameters. This latter property opens up a wide field of practical applications. For convenience we use throughout the appendix the notation of Section 1.2, i.e. y and θ denote sets (scalar, vector or matrix) of observations and parameters respectively.

A5.1.1 Decision Theoretic Justification for Plug-in Estimation.

The Bayesian plug-in estimators (PIE) appear when a modeller is forced to adopt a plug-in rule (Dawid, 1984) in order to choose a likelihood for predicting the next observation. We use the expected LSR as a utility function, i.e. the utility of estimating θ by $\hat{\theta}$ is

$$U(\theta, \hat{\theta}) = \underset{y|\theta}{\text{E}} \log p(y|\hat{\theta}). \quad (\text{A5.1.1})$$

The use of the LSR may be justified by saying that it essentially characterizes the local and proper scoring rules (Bernardo, 1979). A local scoring rule is one where its score only depends on the observation, and can be seen as a generalization of the likelihood principle. A scoring rule is said to be proper if (and only if) the score is maximized by using the actual (according to the modeller) probability density function; in our case it means that if θ is known then the utility function (A5.1.1) is maximized by taking $\hat{\theta} = \theta$.

Therefore, the associated loss function $l(\theta, \hat{\theta}) = \left(\max_{\hat{\theta}} U(\theta, \hat{\theta}) \right) - U(\theta, \hat{\theta})$ is given by,

$$l(\theta, \hat{\theta}) = \underset{y|\theta}{\text{E}} \log \frac{p(y|\theta)}{p(y|\hat{\theta})}, \quad (\text{A5.1.2})$$

and the usual constraints $l(\theta, \hat{\theta}) \geq 0$ and $l(\theta, \theta) = 0$ hold. This loss function is the Kullback and Liebler (1951) directed divergence of the estimated likelihood from the actual likelihood. In addition, it is easily interpreted as the expected weight of evidence - the (natural) logarithm of the Bayes' factor - in favor of a model using the actual likelihood as compared with a model using a estimated likelihood. A motivation for considering expected weight of evidence may be found in Good (1983).

The PIE follows from (A5.1.2) after solving the usual optimization problem, namely,

$$\min_{\hat{\theta}} \underset{\theta}{\text{E}} l(\theta, \hat{\theta}) = \min_{\hat{\theta}} \underset{\theta}{\text{E}} \underset{y|\theta}{\text{E}} \log \frac{p(y|\theta)}{p(y|\hat{\theta})}. \quad (\text{A5.1.3})$$

The invariance of PIE's under one-to-one parametric transformations is apparent from (A5.1.3) since the likelihood itself is invariant. Incidentally it is, of course, this intrinsic property of the likelihood that also makes the maximum likelihood estimators invariant. A related property of PIE's is their invariance

under sufficient transformations of y , in relation to θ . This latter property is easy to verify with the aid of the Fisher-Neyman factorization theorem.

The expected value of the perfect information (EVPI) given by (A5.1.3) can be rewritten as,

$$\min_{\hat{\theta}} \mathbb{E} l(\theta, \hat{\theta}) = \mathbb{E}_{\theta} \mathbb{E}_{y|\theta} \log \frac{p(y|\theta)}{p(\theta)} + \min_{\hat{\theta}} \mathbb{E}_y \log \frac{p(y)}{p(y|\hat{\theta})}. \quad (\text{A5.1.4})$$

Therefore, the problem (A5.1.3) is equivalent to minimizing the Kullback-Liebler directed divergence of the estimated likelihood from the predictive density. This latter problem has been studied by Amaral and Dunsmore (1980). Thus the point estimators found there are PIE's; although their emphasis is in the use of the estimated likelihood as a simple alternative to the predictive density, and no attention is paid to the point estimators as such. Also, it is apparent from (A5.1.4) that the utility function (A5.1.1) is extreme, in the sense that its EVPI vanishes if and only if y and θ are independent.

A5.1.2 The Relationship with the Information Measure.

The first term in the right-hand side of (A5.1.4) denoted by $I(\theta, y)$ can be rewritten as,

$$I(\theta, y) = \mathbb{E}_{\theta} \mathbb{E}_{y|\theta} \log \frac{p(y|\theta)}{p(y)} = \mathbb{E}_{\theta, y} \log \frac{p(\theta, y)}{p(\theta)p(y)} = \mathbb{E}_y \mathbb{E}_{\theta|y} \log \frac{p(\theta|y)}{p(\theta)}. \quad (\text{A5.1.5})$$

This quantity is known as the expected information about θ provided by y (Lindley, 1956), or simply, in view of its symmetry, as the information between θ and y (Pugachev, 1965, Chapter 7). Note that $I(\theta, y) \geq 0$ and the equality holds only if y and θ are independent since the LSR is proper. Moreover, if the statistic s is a function of y then $I(y, \theta) = I(s, \theta) + \mathbb{E}_y \mathbb{E}_{\theta|y} \log \frac{p(\theta|y)}{p(\theta|s)}$, and using again the fact that the LSR is proper it is clear that $I(y, \theta) \geq I(s, \theta)$. The equality holds only if s is a sufficient statistic; see also Kullback and Liebler (1951). Of course, the symmetric result for "sufficient" parametric transformations is valid as well.

The quantity $I(y, \theta)$ can be interpreted via the decision theoretic approach employed in the formulation of the PIE's; suppose that the same LSR as in (A5.1.1) is employed as a utility function but the constraint of using a member of the likelihood family for predictive purposes is no longer imposed, instead, any density $q(y)$ can be used, i.e.

$$U(\theta, q) = \mathbb{E}_{y|\theta} \log q(y). \quad (\text{A5.1.6})$$

Then, the principle of maximizing the expected utility leads to,

$$\max_q \mathbb{E}_y \log q(y), \quad \text{s.t. } q(y) \text{ is a density function.} \quad (\text{A5.1.7})$$

We obtain, using the fact that the LSR is proper, the following reassuring result: the predictive density $p(y) = \int_{\theta} p(y|\theta) p(\theta) d\theta$ is the optimal predictive density. Moreover, the corresponding loss function to (A5.1.6) is

$$l(\theta, q) = \mathbb{E}_{y|\theta} \log p(y|\theta) - \mathbb{E}_{y|\theta} \log q(y) = \mathbb{E}_{y|\theta} \log \frac{p(y|\theta)}{q(y)}. \quad (\text{A5.1.8})$$

The information between y and θ is, in this context, the expected value of the perfect information, i.e.

$$I(y, \theta) = \mathbb{E}_{\theta} l(\theta, p). \quad (\text{A5.1.9})$$

A related motivation for considering $I(y, \theta)$ as a measure of information can be found in Barlow (1985). In view of (A5.1.9) we can interpret the second term in the right-hand side of (A5.1.4) as the expected net loss due to estimation (ENLE) and use it as a companion measure of net risk of the PIE's. Note that the ENLE is necessarily non-negative.

The fact that the LSR is proper is used throughout this Appendix and in Section 3.4. This central result can be easily derived following Pugachev (1965, Chapter 7). Let y be a random variable which is distributed according to the density function $p(y)$ and let $q(y)$ be another density function defined on Y , then the well-known inequality $\log x \leq x - 1$ (with equality only if $x = 1$) implies, for $x = \frac{q(y)}{p(y)}$, that $\mathbb{E}_y \log \frac{q(y)}{p(y)} \leq 0$ (with equality only if $q(y) = p(y)$), and the result follows immediately.

A5.1.3 Standard Static Univariate Models.

The ideas expressed previously are illustrated by looking at some standard static univariate models (using the usual conjugate priors).

(a) Exponential Model

likelihood	:	$p(y \lambda) = \lambda \exp(-\lambda y) \quad (y > 0)$
prior	:	$p(\lambda) = \frac{\beta^{\alpha}}{(\alpha-1)!} \lambda^{\alpha-1} \exp(-\beta\lambda) \quad (\lambda > 0)$
utility	:	$U(\lambda, \hat{\lambda}) = -\frac{\hat{\lambda}}{\lambda} + \log \hat{\lambda}$
loss	:	$l(\lambda, \hat{\lambda}) = \left(\frac{\hat{\lambda}}{\lambda} - 1 \right) + \log \frac{\lambda}{\hat{\lambda}}$
optimum	:	$\hat{\lambda} = \frac{\alpha-1}{\beta}$
EVPI	:	$E_l(\lambda, \hat{\lambda}) = \delta(\alpha) - \log(\alpha - 1)$
information	:	$I(y, \lambda) = \frac{1}{\alpha} + \delta(\alpha) - \log \alpha$
ENLE	:	$E_n(\lambda, \hat{\lambda}) = \log\left(\frac{\alpha}{\alpha-1}\right) - \frac{1}{\alpha}$.

(b) Poisson

likelihood	:	$p(y \lambda) = \frac{\lambda^y}{y!} \exp(-\lambda) \quad (y = 0, 1, \dots)$
prior	:	$p(\lambda) = \frac{\beta^{\alpha}}{(\alpha-1)!} \lambda^{\alpha-1} \exp(-\beta\lambda) \quad (\lambda > 0)$
utility	:	$U(\lambda, \hat{\lambda}) \approx \lambda \log(\hat{\lambda}) - \hat{\lambda} - E \log y!$
loss	:	$l(\lambda, \hat{\lambda}) = \lambda \left(\left(\frac{\hat{\lambda}}{\lambda} - 1 \right) + \log \frac{\lambda}{\hat{\lambda}} \right)$
optimum	:	$\hat{\lambda} = \frac{\alpha}{\beta}$
EVPI	:	$E_l(\lambda, \hat{\lambda}) \approx \frac{\alpha+1}{\beta} (\delta(\alpha+1) - \log \beta) - \frac{\alpha}{\beta} \log \frac{\alpha}{\beta}$.

(c) Bernoulli

$$\begin{aligned}
 \text{likelihood} & : p(y|\theta) = \theta^y (1-\theta)^{1-y} \quad (y = 0, 1) \\
 \text{prior} & : p(\theta) = \frac{(\alpha+\beta-1)}{(\alpha-1)!(\beta-1)!} \theta^{\alpha-1} (1-\theta)^{\beta-1} \quad (0 < \theta < 1) \\
 \text{utility} & : U(\theta, \hat{\theta}) = \theta \log(\hat{\theta}) + (1-\theta) \log(1-\hat{\theta}) \\
 \text{loss} & : l(\theta, \hat{\theta}) = \theta \log\left(\frac{\theta}{\hat{\theta}}\right) + (1-\theta) \log\left(\frac{(1-\theta)}{(1-\hat{\theta})}\right) \\
 \text{optimum} & : \hat{\theta} = \frac{\alpha}{\alpha+\beta} \\
 \text{EVPI} & : E_{\theta} l(\theta, \hat{\theta}) = \frac{\alpha}{(\alpha+\beta)} \left(\delta(\alpha+1) - \delta(\alpha+\beta+1) - \log \frac{\alpha}{(\alpha+\beta)} \right) \\
 & \quad + \frac{\beta}{(\alpha+\beta)} \left(\delta(\beta+1) + \delta(\alpha+\beta+1) - \log \frac{\beta}{(\alpha+\beta)} \right).
 \end{aligned}$$

(d) Pascal

$$\begin{aligned}
 \text{likelihood} & : p(y|\theta) = \frac{(\alpha+\beta-1)}{\binom{\alpha+\beta-1}{y}} \theta^y (1-\theta)^{\alpha+\beta-y-1} \quad (y = 0, 1, \dots) \\
 \text{prior} & : p(\theta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha-1)!(\beta-1)!} \theta^{\alpha-1} (1-\theta)^{\beta-1} \quad (0 < \theta < 1) \\
 \text{utility} & : U(\theta, \hat{\theta}) = \frac{(1-\theta)}{\theta} \log(1-\hat{\theta}) + \log(\hat{\theta}) \\
 \text{loss} & : l(\lambda, \hat{\lambda}) = \frac{(1-\theta)}{\theta} \log\left(\frac{(1-\theta)}{(1-\hat{\theta})}\right) + \log\left(\frac{\theta}{\hat{\theta}}\right) \\
 \text{optimum} & : \hat{\theta} = \frac{\alpha-1}{\alpha+\beta-1} \\
 \text{EVPI} & : E_{\theta} l(\theta, \hat{\theta}) = \frac{(\alpha+\beta-1)}{(\alpha-1)} \left(\delta(\beta) - \delta(\alpha+\beta-1) - \log \frac{\beta}{\alpha+\beta-1} \right) \\
 & \quad + \left(\delta(\alpha) - \delta(\alpha+\beta) - \log \left(\frac{(\alpha-1)}{(\alpha+\beta-1)} \right) \right) \\
 & \quad - \left(\delta(\beta) - \delta(\alpha+\beta) - \log \frac{\beta}{\alpha+\beta-1} \right).
 \end{aligned}$$

(e) Normal

$$\begin{aligned}
 \text{likelihood} & : y|\theta, \sigma^2 \sim N(\mu, \sigma^2) \quad (-\infty < y < \infty) \\
 \text{prior} & : \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} \sim \text{NG}^{-1}(m, cs, d) \quad (-\infty < \theta < \infty, \quad \sigma^2 > 0) \\
 \text{utility} & : U\left(\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}, \begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix}\right) = -\frac{1}{2} \left(\frac{\sigma^2 + (\mu - \hat{\mu})^2}{\hat{\sigma}^2} + \log(2\pi) + \log \sigma^2 \right) \\
 \text{loss} & : l\left(\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}, \begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix}\right) = \frac{1}{2} \left(\frac{(\mu - \hat{\mu})^2}{\hat{\sigma}^2} + \left(\frac{\sigma^2}{\hat{\sigma}^2} - 1 \right) + \log \frac{\sigma^2}{\hat{\sigma}^2} \right) \\
 \text{optimum} & : \hat{\mu} = m, \quad \hat{\sigma}^2 = (1+c) \frac{S}{(d-2)} \\
 \text{EVPI} & : E_{\theta, \sigma^2} l\left(\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}, \begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix}\right) = \frac{1}{2} (\log(1+c) + \delta(\frac{1}{2}d) - \log(\frac{1}{2}d-1)) \\
 \text{information} & : I\left(y, \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}\right) = \frac{1}{2} \log(1+c) - \frac{1}{2} + \log \left(\frac{(\frac{1}{2}d-1)!}{(\frac{1}{2}d-\frac{1}{2})!} \right) \\
 & \quad + (\frac{1}{2}d + \frac{1}{2}) \delta(\frac{1}{2}d + \frac{1}{2}) - \frac{1}{2} d \delta(\frac{1}{2}d).
 \end{aligned}$$

From these examples it can be seen that the utility function (A4.1) induces a sensible loss function which automatically considers the role of the estimators of the parameters for predictive purposes. Moreover, in some cases it is possible to quantify the ENLE. For instance, it can be shown that $I(y, \lambda) = \frac{1}{\alpha} + \delta(\alpha) - \log \alpha$ for the exponential model and this allows us to make a comparison of the ENLE in relation with the EVPI and $I(y, \lambda)$, depending on the values of the shape parameter α as is shown in Table A5.1.1. It is reassuring that the ENLE becomes rapidly negligible relative to the EVPI for α 's as small as 10. In other cases, as in the Poisson model, it seems that there is no way of obtaining the value of the information other than by approximations, but then such information is bounded by the EVPI, and for informative priors the EVPI itself can be used as an approximation of the information.

The above discussion and the decision theoretic justification of the PIE emphasize their usefulness in prediction. Nevertheless, we are not suggesting an indiscriminate use of these estimators, to do so would be as naive as to always suggest the use of the expected information as expected utility. However, we strongly recommend them as first refusal estimators for (functions of) the parameters, i.e. to use them unless there is a better estimation procedure available. Finally, it is important to note that there are statistical models for which the PIE's may not exist, for instance, when the support of the predictive density does not coincide with the support of any likelihood chosen as possible approximand.

alpha	EVPI	INFO	ENLE
1.5	1.3963	0.9644	0.431946
2.0	0.9228	0.7296	0.193147
2.5	0.6977	0.5869	0.110826
3.0	0.5630	0.4908	0.072132
3.5	0.4726	0.4218	0.050758
4.0	0.4075	0.3698	0.037682
4.5	0.3583	0.3292	0.029092
5.0	0.3198	0.2967	0.023144
6.0	0.2633	0.2477	0.015655
7.0	0.2239	0.2126	0.011294
8.0	0.1947	0.1862	0.008531
9.0	0.1723	0.1656	0.006672
10.0	0.1545	0.1432	0.005361
20.0	0.0761	0.0748	0.001293
30.0	0.0505	0.0499	0.000568
40.0	0.0378	0.0374	0.000318
50.0	0.0302	0.0300	0.000203
100.0	0.0150	0.0150	0.000050
150.0	0.0100	0.0100	0.000022
200.0	0.0075	0.0075	0.000013
250.0	0.0060	0.0060	0.000008
300.0	0.0050	0.0050	0.000006
350.0	0.0043	0.0043	0.000004
400.0	0.0038	0.0037	0.000003
450.0	0.0033	0.0033	0.000002
500.0	0.0030	0.0030	0.000002

Table A5.1.1 Values of EVPI, information and ENLE as functions of the shape parameter for the exponential model.

APPENDIX A5.2.

ENTROPY AND INFORMATION OF USEFUL MATRIX-VARIATE DISTRIBUTIONS.

The following results are useful for deriving formulas concerning the information and entropy of some matrix-variate distributions. The entropy of a random variable Y is defined by $H(y) = -\mathbb{E} \log_p(p(y))$.

Let $\Theta \sim N(M, C, \Sigma)$ (given Σ), then

$$\mathbb{E}_{\Theta|\Sigma} (\Theta - L)' A (\Theta - L) = (M - L)' A (M - L) + \Sigma \text{tr}(AC) \quad (\text{A5.2.1a})$$

where A is positive semidefinite. From (A3.2.5) and (A3.2.7b) we have,

$$\begin{aligned} \mathbb{E}_{\Theta|\Sigma} (\Theta - L)' A (\Theta - L) &= \mathbb{E}_{\Theta|\Sigma} ((\Theta - M) + (M - L))' A ((\Theta - M) + (M - L)) \\ &= (M - L)' A (M - L) + \mathbb{E}_{\Theta|\Sigma} (B(\Theta - M))' (B(\Theta - M)) \\ &= (M - L)' A (M - L) + \Sigma \text{tr}(AC), \end{aligned}$$

where $B'B$ is the Cholesky decomposition of A ; see Appendix 4.1. If in addition $\Sigma \sim W^{-1}(S, d)$, then

$$\mathbb{E}_{\Theta} (\Theta - L)' A (\Theta - L) = (M - L)' A (M - L) + \frac{S}{(d-2)} \text{tr}(AC). \quad (\text{A5.2.1b})$$

This follows directly from (A5.2.1a) and (A3.2.14). Note that in particular, formula (A5.2.1b) holds for $\Theta \sim T(M, C, S, d)$.

Let $\Sigma \sim W^{-1}(S, d)$, then

$$\mathbb{E}_{\Sigma} \Sigma^{-1} = (d + q - 1) S^{-1} \quad (\text{A5.2.2a})$$

and

$$\mathbb{E}_{\Sigma} \log |\Sigma| = \log(|\frac{1}{2} S|) - \sum_{j=1}^q \delta(\frac{1}{2}(\nu - j + 1)). \quad (\text{A5.2.2b})$$

Note that (A5.2.2a) is just the well-known formula for the mean of a Wishart random matrix. The result (A5.2.2) may be verified as a consequence of a more general result. Let us consider the problem of approximating the distribution of a positive definite random matrix Σ by an inverted-Wishart using the LSR (see Appendix A5.1) as the utility function. The optimization problem is,

$$\max_{\hat{S}, \hat{d}} \mathbb{E}_{\Sigma} \log p(\Sigma), \quad (\text{A5.2.3})$$

where the expectation is taken over Σ (the target random matrix) and $p(\Sigma)$ stands for the density function of the inverted-Wishart distribution $W^{-1}(\hat{S}, \hat{d})$. Following the usual (differentiating) procedure we obtain the optimal conditions for \hat{S} and \hat{d} ,

$$\mathbb{E}_{\Sigma} \Sigma^{-1} = (\hat{d} + q - 1) \hat{S}^{-1} \quad (\text{A5.2.4a})$$

and

$$\mathbb{E}_{\Sigma} \log |\Sigma| = \log |\frac{1}{2} \hat{\Sigma}| - \sum_{j=1}^q \delta(\frac{1}{2}(\hat{d} + q - j)). \quad (\text{A5.2.4b})$$

Therefore, for $\Sigma \sim W^{-1}(S, d)$ the formula (A5.2.4) becomes (A5.2.2) since the LSR is proper.

A5.2.1 Matrix-normal (non-singular).

The entropy of a matrix-normal random variable $\theta \sim N(M, C, \Sigma)$ is given by

$$H(\Theta) = \frac{1}{2}pq(1 + \log(2\pi)) + \frac{1}{2}q \log(|C|) + \frac{1}{2}p \log(|\Sigma|). \quad (\text{A5.2.5})$$

This formula follows easily from (A3.2.10) and (A5.2.5).

Let $\Theta = \begin{bmatrix} \Theta_1 \\ \Theta_2 \end{bmatrix}$ where Θ is distributed as before. Then, the information between Θ_1 and Θ_2 is given by

$$I(\Theta_1, \Theta_2) = \frac{1}{2}q \log\left(\frac{|C_{11}| |C_{22}|}{|C|}\right) = \frac{1}{2}q \log\left(\frac{|C_{22}|}{|C_{22|1}|}\right) = \frac{1}{2}q \log\left(\frac{|C_{11}|}{|C_{11|2}|}\right), \quad (\text{A5.2.6})$$

where $C_{22|1}$ and $C_{11|2}$ are defined as in (A3.2.8). Of course, an obvious symmetric formula holds for subsets of columns. The result (A5.2.6) follows easily from (A5.2.5) by noticing that $I(\Theta_1, \Theta_2) = H(\Theta_1) + H(\Theta_2) - H(\Theta_1, \Theta_2)$. Moreover, it is not difficult to see that the first equality in (A5.2.6) is still valid if the matrix variances are replaced by their corresponding correlation matrices.

A5.2.2 Inverted-Wishart.

The entropy of an inverted-Wishart random matrix $\Sigma \sim W^{-1}(S, d)$, is given by,

$$H(\Sigma) = \frac{1}{2}q\nu + \frac{q(q-1)}{4} \log(\pi) + \sum_{j=1}^q \log\left(\left(\frac{1}{2}(\nu-j-1)\right)!\right) + \frac{1}{2}(q+1) \log\left(\frac{1}{2}|S|\right) - \frac{1}{2}(d+q) \sum_{j=1}^q \delta\left(\frac{1}{2}(\nu-i+1)\right), \quad (\text{A5.2.7})$$

where $\nu = d + q - 1$. This formula follows directly from (A3.2.15) and (A5.2.2).

A5.2.3 Matrix-normal Inverted-Wishart.

The information between Θ and Σ , where $\begin{bmatrix} \Theta \\ \Sigma \end{bmatrix} \sim NW^{-1}(M, C, S, d)$ is given by,

$$I(\Theta, \Sigma) = -\frac{1}{2}qp + \sum_{j=1}^q \log\left(\frac{\left(\frac{1}{2}(\nu-j-1)\right)!}{\left(\frac{1}{2}(\nu+p-j-1)\right)!}\right) + \frac{1}{2}(\nu+p) \sum_{j=1}^q \delta\left(\frac{1}{2}(\nu+p-j+1)\right) - \frac{1}{2}\nu \sum_{j=1}^q \delta\left(\frac{1}{2}(\nu-j+1)\right), \quad (\text{A5.2.8})$$

where $\nu = d + q - 1$ as before. After using (A5.2.7), (A3.2.19),

$$I(\Theta, \Sigma) = \mathbb{E}_{\Theta} \mathbb{E}_{\Sigma|\Theta} \log p(\Sigma|\Theta) - \mathbb{E}_{\Sigma} \log p(\Sigma) \quad (\text{A5.2.9})$$

and

$$\mathbb{E}_{\Theta} \log \left| \frac{1}{2}(S + (\Theta - M)'C^{-1}(\Theta - M)) \right| = \log \left| \frac{1}{2}S \right| + \sum_{j=1}^q (\delta\left(\frac{1}{2}(\nu+p-j+1)\right) - \delta\left(\frac{1}{2}(\nu-j+1)\right)), \quad (\text{A5.2.10})$$

the derivation of (A5.2.8) is reduced to an exercise of algebra. Formula (A5.2.10) follows from (A5.2.2b), (A3.2.19) and $\mathbb{E}_{\Sigma} \log |\Sigma| = \mathbb{E}_{\Theta} \mathbb{E}_{\Sigma|\Theta} \log |\Sigma|$.

CHAPTER 6
DYNAMIC RECURSIVE MODEL
AND
DYNAMIC SCALE VARIANCE

Two tractable dynamic multivariate models are formulated in this chapter. The first uses the ideas of the econometric recursive models in order to combine dynamic tractable models into a more complex but still tractable dynamic model. The second relies on the discount method for simulating a DLMR with a dynamic scale variance. Two examples illustrate the methods proposed, one with artificially generated data and the other using exchange rate data.

Regardless of how easy the sequential implementation of a model might be, there are always inferential problems, such as finding the mean of a non-trivial function of a parameter, which demand a numerical method for their solution. In Appendix A6.1 the simulation technique is reviewed briefly and a Bayesian justification is discussed also. The distribution of swept matrices plays an important role in understanding the material of Sections 3 and 4. This distribution theory is presented in Appendix A6.2.

6.1 DYNAMIC RECURSIVE MODEL.

The standard Bayesian analysis of fully recursive (often called causal) models, in the context of simultaneous equation econometric models, leads to results completely analogous to those of the multiple regression model (Zellner, 1971, p. 250-252). This suggests that a dynamic recursive model may be formulated in order to admit a tractable analysis and implementation. In this section we present a general formulation and analysis of such a model. Then we apply our results to a dynamic linear recursive model. Throughout this section we use the same notation as that of Section 1.2, i.e. y and θ denote a set (scalar, vector or matrix) of observations and parameters respectively

Let us consider, first, a simple model for the time series y_{1t}, y_{2t} . Our observation, evolution and prior information assumptions at time t are:

$$p(y_{1t}, y_{2t} | \theta_{1t}, \theta_{2t}) = p(y_{1t} | \theta_{1t}) p(y_{2t} | \theta_{2t}, y_{1t}), \quad (6.1a)$$

$$p(\theta_{1t}, \theta_{2t} | \theta_{1(t-1)}, \theta_{2(t-1)}) = p(\theta_{1t} | \theta_{1(t-1)}) p(\theta_{2t} | \theta_{2(t-1)}), \quad \text{and} \quad (6.1b)$$

$$p(\theta_{1(t-1)}, \theta_{2(t-1)}) = p(\theta_{1(t-1)}) p(\theta_{2(t-1)}). \quad (6.1c)$$

The observation equation (6.1a) is twofold. Firstly, y_{1t} and θ_{2t} are considered (conditionally) independent given θ_{1t} . Secondly, y_{2t} and θ_{1t} are considered independent given y_{1t} and θ_{2t} . The evolutional assumption (6.1b) is threefold. Firstly, it says that θ_{1t} and θ_{2t} are independent given $\theta_{1(t-1)}$ and $\theta_{2(t-1)}$. Secondly, θ_{1t} and $\theta_{2(t-1)}$ are independent given $\theta_{1(t-1)}$. Thirdly, θ_{2t} and $\theta_{1(t-1)}$ are independent given

$\theta_{2(t-1)}$. The prior information assumption simply says that $\theta_{1(t-1)}$ and $\theta_{2(t-1)}$ are apriori independent. In addition, the usual independence over time is assumed, i.e. the evolution of the parameters is Markovian and present observations are independent of past observations given the present parameters.

It is easy to see that the evolution, prediction and posterior recurrences are given by,

$$p(\theta_{1t}, \theta_{2t}) = p(\theta_{1t}) p(\theta_{2t}), \quad (6.2a)$$

$$p(y_{1t}, y_{2t}) = p(y_{1t}) p(y_{2t}|y_{1t}), \quad \text{and} \quad (6.2b)$$

$$p(\theta_{1t}, \theta_{2t}|y_{1t}, y_{2t}) = p(\theta_{1t}|y_{1t}) p(\theta_{2t}|y_{1t}, y_{2t}). \quad (6.2c)$$

The right-hand side densities in (6.2) are given by the evolution, prediction and posterior recurrence formulas corresponding to the dynamic submodels defined by the right-hand side densities in (6.1), namely

$$p(y_{1t}, |\theta_{1t}), \quad p(\theta_{1t}|\theta_{1(t-1)}), \quad p(\theta_{1(t-1)}), \quad \text{and} \quad (6.3a)$$

$$p(y_{2t}|\theta_{2t}, y_{1t}), \quad p(\theta_{2t}|\theta_{2(t-1)}), \quad p(\theta_{2(t-1)}). \quad (6.3b)$$

Notice that the key feature of the analysis is that the conditional prior independence of the state parameters is preserved by the evolution and posterior recurrence formulas. Although data y_{1t} appears in the posterior for θ_{2t}

6.1.1 Model Formulation.

The dynamic recursive model formulation and analysis are the straightforward extensions of (6.1) and (6.2) to several submodels. For convenience we denote a set of contemporaneous observations (y_{1t}, \dots, y_{mt}) by $y_t^{(m+1)}$, similarly $\theta_t^{(m+1)}$ stands for $(\theta_{1t}, \dots, \theta_{mt})$. The assumptions for the dynamic recursive model are given below.

Observation Equation:

$$p(y_t^{(m+1)}|\theta_t^{(m+1)}) = \prod_{k=1}^m p(y_{kt}|\theta_{kt}, y_t^{(k)}). \quad (6.4a)$$

Evolution Equation:

$$p(\theta_t^{(m+1)}|\theta_{t-1}^{(m+1)}) = \prod_{k=1}^m p(\theta_{kt}|\theta_{k(t-1)}). \quad (6.4b)$$

Prior Information:

$$p(\theta_{t-1}^{(m+1)}) = \prod_{k=1}^m p(\theta_{k(t-1)}). \quad (6.4c)$$

In addition the usual independence over time is assumed.

It is important to note that $p(y_{kt}|\theta_{kt}, y_t^{(k)})$ means only possible dependence upon $\theta_{kt}, y_{1t}, \dots, y_{(k-1)t}$. In particular, hierarchical structures of dependence among observations are allowed; see, for example, the model (6.8) of Section 6.2.

6.1.2 Updating Procedure.

The updating recurrences corresponding to model (6.4) are:

Evolution:

$$p(\theta_t^{(m+1)}) = \prod_{k=1}^m p(\theta_{kt}), \quad (6.5a)$$

where $p(\theta_{kt})$ is as in (6.6a) below.

Prediction:

$$p(y_t^{(m+1)}) = \prod_{k=1}^m p(y_{kt}|y_t^{(k)}), \quad (6.5b)$$

where $p(y_{kt}|y_t^{(k)})$ is as in (6.6b) below.

Posterior:

$$p(\theta_t^{(m+1)}|y_t^{(m+1)}) = \prod_{k=1}^m p(\theta_{kt}|y_t^{(k+1)}), \quad (6.5c)$$

where $p(\theta_{kt}|y_t^{(k+1)})$ is as in (6.6c) below.

These formulas may be verified as follows. From (6.4b) and (6.4c) we have,

$$\begin{aligned} \int_{\theta_{t-1}^{(m+1)}} p(\theta_t^{(m+1)}|\theta_{t-1}^{(m+1)}) p(\theta_{t-1}^{(m+1)}) d\theta_{t-1}^{(m+1)} &= \int_{\theta_{t-1}^{(m+1)}} \prod_{k=1}^m p(\theta_{kt}|\theta_{k(t-1)}) p(\theta_{k(t-1)}) d\theta_{t-1}^{(m+1)} \\ &= \prod_{k=1}^m \int_{\theta_{k(t-1)}} p(\theta_{kt}|\theta_{k(t-1)}) p(\theta_{k(t-1)}) d\theta_{k(t-1)}, \end{aligned}$$

which is formula (6.5a). From (6.4) and (6.5a) we have,

$$\begin{aligned} \int_{\theta_t^{(m+1)}} p(y_t^{(m+1)}|\theta_t^{(m+1)}) p(\theta_t^{(m+1)}) d\theta_t^{(m+1)} &= \int_{\theta_t^{(m+1)}} \prod_{k=1}^m p(y_{kt}|\theta_{kt}, y_t^{(k)}) d\theta_t^{(m+1)} \\ &= \prod_{k=1}^m \int_{\theta_{kt}} p(y_{kt}|\theta_{kt}, y_t^{(k)}) p(\theta_{kt}) d\theta_{kt}, \end{aligned}$$

which is formula (6.5b). Finally, from (6.4a), (6.5a) and (6.5b) we have,

$$\frac{p(y_t^{(m+1)}|\theta_t^{(m+1)}) p(\theta_t^{(m+1)})}{p(y_t^{(m+1)})} = \prod_{k=1}^m \frac{p(y_{kt}|\theta_{kt}, y_t^{(k)}) p(\theta_{kt})}{p(y_{kt}|y_t^{(k)})},$$

which is formula (6.5c).

The prior independence (6.4c) of the parameters is, as before, preserved by the evolution recurrence (6.5a) and the posterior recurrence (6.5c). Furthermore, in view of (6.5), the updating procedure of the dynamic recursive model (6.4) can be decomposed in terms of the updating recurrences of the dynamic submodels defined by the right-hand side densities in (6.4). In other words, the evolution, prediction and posterior updating recurrences,

$$p(\theta_{kt}) = \int_{\theta_{kt}} p(\theta_{kt}|\theta_{k(t-1)}) p(\theta_{k(t-1)}) d\theta_{k(t-1)}, \quad (6.6a)$$

$$p(y_{kt}|y_t^{(k)}) = \int_{\theta_{kt}} p(y_{kt}|\theta_{kt}, y_t^{(k)}) p(\theta_{kt}) d\theta_{kt}, \quad (6.6b)$$

$$p(\theta_{kt}|y_t^{(k+1)}) = \frac{p(y_{kt}|\theta_{kt}, y_t^{(k)}) p(\theta_{kt})}{p(y_{kt}|y_t^{(k)})}, \quad (6.6c)$$

for $k = 1, \dots, m$ corresponding to the submodels whose observational, evolution and prior densities are $p(y_{kt}|\theta_{kt}, y_t^{(k)})$, $p(\theta_{kt}|\theta_{k(t-1)})$, $p(\theta_{k(t-1)})$, can be carried out in parallel. The evolution, prediction and posterior densities of the whole system always can be recovered by means of (6.5a), (6.5b) and (6.5c). Therefore, if the updating recurrences (6.6) of the submodels are tractable the whole system has a tractable implementation (depending, of course, on the number of submodels considered).

6.1.3 Dynamic Linear Recursive Model.

A dynamic linear recursive models is simply a dynamic recursive model (6.4), such that each submodel is a DLMR. Therefore, according to the previous results, its implementation is straightforward; it is only necessary to implement in parallel the defining submodels. This can be accomplished by means of any of the algorithms discussed in Chapter 4, in particular, the implementation via the sweep operator is recommended in view of its simplicity and versatility.

Let us consider an example in which all submodels are univariate DLM's; a dynamic version of the fully recursive model (Zellner, 1971, p. 250). The assumptions of the model are as follows.

Observation:

$$y_{kt} = \underline{z}'_{kt} \underline{\delta}_{kt} + e_{kt}, \quad e_{kt} \sim N(0, v_{kt} \sigma_k^2), \quad k = 1, \dots, m, \quad (6.7a)$$

where

$$\begin{aligned} \underline{z}'_{kt} &= (\underline{y}_t^{(k)}', \underline{x}'_{kt}), & \underline{y}_t^{(k)} &= (y_{1t}, \dots, y_{(k-1)t}), & \underline{x}'_{kt} &= (x_{1t}, \dots, x_{pt}), \\ \underline{\delta}'_{kt} &= (\underline{\gamma}_k^{(k)}, \underline{\theta}'_{kt}), & \underline{\gamma}_k^{(k)}' &= (\gamma_{k1t}, \dots, \gamma_{k(k-1)t}), & \underline{\theta}_{kt} &= (\theta_{1t}, \dots, \theta_{pt}), \end{aligned}$$

and the errors e_{kt} ($k = 1, \dots, m$) are independent given $\underline{\sigma}^2 = (\sigma_1^2, \dots, \sigma_k^2)$.

Evolution:

$$\underline{\delta}_{kt} = G_{kt} \underline{\delta}_{k(t-1)} + \underline{f}_{kt}, \quad \underline{f}_{kt} \sim N(0, \sigma_k^2 W_{kt}), \quad k = 1, \dots, m, \quad (6.7b)$$

where the noises \underline{f}_{kt} ($k = 1, \dots, m$) are independent given $\underline{\sigma}^2$.

Prior:

$$\underline{\delta}_{k(t-1)} \sim N(\underline{m}_{k(t-1)}, C_{k(t-1)}), \quad \sigma_{kt}^2 \sim \Gamma^{-1}(\frac{1}{2}d_{k(t-1)}, \frac{1}{2}s_{k(t-1)}), \quad k = 1, \dots, m, \quad (6.7c)$$

where $\underline{f}_{k(t-1)}$ ($k = 1, \dots, m$) are independent given $\underline{\sigma}^2$ and σ_{kt}^2 are also mutually independent.

In the context of simultaneous equation econometric models the parameters $\underline{\gamma}_{kt}^{(k)}$ are referred to as the coefficients of the endogenous variables. In particular, taking $G_{kt} = I$, $v_{kt} = 1$ and $W_{kt} = O$ for all k and t the model becomes the standard, static, fully recursive model. The dynamic model (6.7)

has the usual advantages over its static counterpart, for instance, taking $G_{kt} = \begin{bmatrix} I & O \\ O & H_{kt} \end{bmatrix}$ and W_{kt} non-zero in the evolution equations (6.7b) allows for steady time varying endogenous coefficients and provides a flexible model.

The analysis for a dynamic linear model is, in many respects, analogous to that of the DLMR, for instance, long-term predictions can be achieved following the corresponding procedure (see Section 3.2) for each submodel. However, there are certain aspects that require special attention, e.g. marginal predictive densities and missing observations. Although there is a closed form for the joint predictive density (6.5b), this is not the case in general for the marginal predictive densities. Missing observations can be handled in a tractable way when they conform to the partial order induced by the dependency among the observables (see formula 6.4a); this aspect is discussed in Section 6.2.

The dynamic linear recursive model provides a powerful environment for modelling multivariate time series and yet its analysis is surprisingly simple. Nevertheless, the analysis depends on the partial order induced by the dependency among the observables, and therefore in a real situation this order has to be chosen with great care. For this purpose, it is useful to have in mind the particular characteristics of the time series, e.g. considerations regarding causality between the observables may suggest a suitable order.

6.2 EXAMPLE: HIERARCHICAL MISSING OBSERVATIONS.

The densities of the defining submodels in the right-hand side of (6.4a) induce a partial order between the y_{kt} 's of the dynamic recursive model as follows. We say that $y_{\beta t}$ is a successor of $y_{\alpha t}$ if and only if $y_{\alpha t}$ appears in the density of $y_{\beta t}$. Furthermore, $y_{\beta t}$ is a descendant of $y_{\alpha t}$ either directly (if $y_{\beta t}$ is a successor of $y_{\alpha t}$) or by transitivity. This construction is completely analogous to that of influence diagrams, see for example Barlow (1986) and references therein.

Let us suppose that for any time t , the missing observations conform to the above partial order, i.e. if an observation of a given submodel is missing then necessarily all its descendants are also missing. With these assumptions and keeping in mind the updating procedure for the dynamic recursive model it follows that the missing observations can be handled via the submodels in the usual trivial way; the corresponding posteriors are merely equal to their priors. This is not just an academic result; a major reason for considering a dynamic recursive model is when the data of a time series of interest is sparse, but the data of a related time series is not. Thus, the information can be transferred from the latter time series to the first by means of a dynamic recursive model. The following simple example illustrates this procedure.

6.2.1 Model.

For convenience we entertain a static linear recursive model, its dynamic extension is easily appreciated.

Observation:

$$y_{1t} = \theta_{11} + \theta_{12}t + e_{1t}, \quad e_{1t} \sim N(0, \sigma_1^2), \quad (6.8a)$$

$$y_{2t} = \gamma_{21}y_{1t} + \theta_{21} + e_{2t}, \quad e_{2t} \sim N(0, \sigma_2^2), \quad (6.8b)$$

$$y_{3t} = \gamma_{31}y_{1t} + \theta_{31} + e_{3t}, \quad e_{3t} \sim N(0, \sigma_3^2). \quad (6.8c)$$

Evolution: none.

Prior: vague (see Subsection 4.1.4 and Section 7.1).

The above model represents a time series y_{1t} with a linear trend, and two time series y_{2t} and y_{3t} related with the former in a familiar, linear form. The partial order induced by 6.8 is very simple: y_{1t} has two descendants, y_{2t} and y_{3t} . Furthermore, we assume that the missing data hierarchy conforms to this order, i.e. y_{2t} and/or y_{3t} may be missing at any given time, but we cannot observe either y_{2t} or y_{3t} when y_{1t} is missing.

6.2.2 Artificial Data.

A random sample of model (6.8) was simulated for $t = 1, 2, \dots, 30$ using the setting,

$$\theta_{11} = 2, \quad \theta_{12} = .2, \quad \sigma_1^2 = .75, \quad (6.9a)$$

$$\gamma_{21} = 1.5, \quad \theta_{21} = 1, \quad \sigma_2^2 = 1.5, \quad (6.9b)$$

$$\gamma_{31} = 2, \quad \theta_{31} = 2, \quad \sigma_3^2 = 2. \quad (6.9c)$$

Moreover, observations of the second and third time series were missing with probabilities $\frac{2}{3}$ and $\frac{1}{2}$ respectively. This data is shown in Table 6.1

6.2.3 Analysis.

The observational equation (6.8) can be rewritten in the so-called structural form,

$$\underline{y}'_t \Gamma = \underline{x}'_t \Theta + \underline{\epsilon}'_t, \quad \underline{\epsilon}'_t \sim N(\underline{0}, \Delta^2), \quad (6.10)$$

where

$$\begin{aligned} \underline{y}'_t &= [y_{1t}, y_{2t}, y_{3t}], & \Gamma &= \begin{bmatrix} 1 & -\gamma_{21} & -\gamma_{31} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, & \underline{x}'_t &= [1, t], \\ \Theta &= \begin{bmatrix} \theta_{11} & \theta_{21} & \theta_{31} \\ \theta_{12} & 0 & 0 \end{bmatrix}, & \underline{\epsilon}'_t &= [e_{1t}, e_{2t}, e_{3t}] \quad \text{and } \Delta = \text{diag}(\sigma_1, \sigma_2, \sigma_3). \end{aligned}$$

The reduced form (Zellner, 1971) corresponding to (6.10) is,

$$\underline{y}'_t = \underline{x}'_t \Pi + \underline{d}'_t, \quad \underline{d}'_t \sim N(\underline{0}, \Omega), \quad (6.11)$$

γ_1	γ_2	γ_3	t
2.008026	3.79275	6.595217	1
3.504832	0	0	2
3.602766	0	9.598586	3
2.225966	0	8.793272	4
3.758166	0	0	5
3.37933	8.647253	9.350556	6
5.617537	0	12.94539	7
3.775815	0	0	8
3.483832	0	0	9
4.129178	0	9.959234	10
3.719503	7.489043	8.167546	11
4.122363	0	0	12
3.730811	0	7.325654	13
5.058289	8.034336	15.01	14
4.8941	7.150683	10.37654	15
3.484249	5.307671	0	16
6.570957	14.38003	15.75046	17
3.939957	0	10.48378	18
5.769405	0	16.43938	19
5.361767	0	11.6875	20
7.571926	0	0	21
6.619864	0	17.3816	22
8.041594	0	0	23
7.023859	0	0	24
5.567156	0	0	25
6.151823	11.18707	0	26
6.581729	0	15.70809	27
5.864465	0	0	28
6.795583	0	14.08167	29
8.227899	13.63449	18.62876	30

Table 6.1 Artificial Data

where

$$\Pi = \Theta\Gamma^{-1}, \quad \Omega = \Gamma^{-1'}\Delta^2\Gamma^{-1} \quad \text{and} \quad \Gamma^{-1} = \begin{bmatrix} 1 & \gamma_{12} & \gamma_{31} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Equation (6.11) is merely a convenient representation of (6.4a). Notice that the joint distribution of Π and Ω is implicitly given by the joint distribution of Θ , Γ^{-1} and Δ^2 . Hence, all the necessary information for making inferences about Π and/or Ω given the data is contained in the posterior distributions of the parameters corresponding to each submodel.

Suppose, for instance, that we are interested in the correlations ρ_{12} , ρ_{13} , ρ_{23} associated with Ω . Then, it follows from (6.11) that,

$$\rho_{12} = \frac{\gamma_{21}}{(\gamma_{21}^2 + \alpha_2)^{\frac{1}{2}}}, \quad \rho_{13} = \frac{\gamma_{31}}{(\gamma_{31}^2 + \alpha_3)^{\frac{1}{2}}} \quad \text{and} \quad \rho_{23} = \rho_{12}\rho_{13}, \quad (6.12)$$

where $\alpha_2 = \frac{\sigma_2^2}{\sigma_1^2}$ and $\alpha_3 = \frac{\sigma_3^2}{\sigma_1^2}$. Therefore, in principle, all that we need is the joint posterior density of $\gamma_{21}, \gamma_{31}, \sigma_1^2, \sigma_2^2$ and σ_3^2 which is given by,

$$\sigma_1^2 \sim \Gamma^{-1}\left(\frac{1}{2}d_1, \frac{1}{2}s_1\right), \quad (6.13a)$$

$$\gamma_{21} \sim N(m_{21}, c_{21}\sigma_2^2), \quad \sigma_2^2 \sim \Gamma^{-1}\left(\frac{1}{2}d_2, \frac{1}{2}s_2\right), \quad (6.13b)$$

$$\gamma_{31} \sim N(m_{31}, c_{31}\sigma_3^2), \quad \sigma_3^2 \sim \Gamma^{-1}\left(\frac{1}{2}d_3, \frac{1}{2}s_3\right), \quad (6.13c)$$

where $s_1 = 24.88, d_1 = 30, s_2 = 19.37, d_2 = 9, s_3 = 38.63, d_3 = 18, m_{21} = 1.679, c_{21} = .0339, m_{31} = 1.992, c_{31} = .0206$. Unfortunately, the posterior distributions of the correlations are extremely complex; they cannot be expressed in closed form. Nevertheless, plug-in estimates using unit correction factors are easily obtainable, $\hat{\rho}_{12} = .721, \hat{\rho}_{13} = .778$ and $\hat{\rho}_{23} = .561$ (the actual values are $\rho_{12} = .728, \rho_{13} = .775$ and $\rho_{23} = .564$).

Furthermore, simulation techniques may be employed for approximating other quantities of interest (see Appendix A6.1) such as the mean and the probability that a correlation lies in a neighborhood around its plug-in estimate, etc. We report the results of a simulation of size 2500. This assures an error standard deviation of less than one percent when approximating probabilities; see (A6.1.3). The approximated means were $E \rho_{12} = .697, E \rho_{13} = .766$ and $E \rho_{23} = .536$ and the approximated probabilities were $P(|\rho_{12} - \hat{\rho}_{12}| < .1) = .3588, P(|\rho_{13} - \hat{\rho}_{13}| < .1) = .508$ and $P(|\rho_{23} - \hat{\rho}_{23}| < .1) = .3388$. As a reference, the following quantities were also approximated, $E \alpha_2 = 3.286(3.336), E \alpha_3 = 2.899(2.911), P(.549 < \alpha_2(s_1/d_1)(s_2/d_2) < 1.958) = .748(.75)$ and $P(.617 < \alpha_3(s_1/d_1)(s_3/d_3) < 1.654) = .7548(.75)$ (the theoretical values appear between parenthesis).

These simulation results show that the plug-in based estimates were very accurate compared to the simulated means. In addition, the example illustrates how the simulation techniques of Appendix A6.1 can be combined with plug-in estimates in order to provide not only point estimates, but also good approximations of the concentration of the probability around them.

6.3 DYNAMIC SCALE VARIANCE.

In this section we consider multivariate dynamic linear model with varying observational variance-covariance structure.

The simple bivariate DWMR model,

$$[y_1, y_2]_t = [\theta_1, \theta_2]_t + [e_1, e_2]_t, \quad [e_1, e_2]_t \sim N((0, 0), \Sigma), \quad (6.14a)$$

$$[\theta_1, \theta_2]_t = g[\theta_1, \theta_2]_{t-1} + [f_1, f_2]_t, \quad [f_1, f_2]_t \sim N((0, 0), w\Sigma), \quad (6.14b)$$

$$\begin{bmatrix} \theta_{1(t-1)} & \theta_{2(t-1)} \\ \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \sim NW^{-1}([m_1, m_2]_{t-1}, c_{t-1}S_{t-1}, d_{t-1}). \quad (6.14c)$$

can be decomposed into the following univariate submodels:

Submodel 1:

$$y_{1t} = \theta_{1t} + e_{1t}, \quad e_{1t} \sim N(0, \sigma_{11}), \quad (6.15a)$$

$$\theta_{1t} = g\theta_{1(t-1)} + f_{1t}, \quad f_{1t} \sim N(0, w\sigma_{11}), \quad (6.15b)$$

$$\begin{bmatrix} \theta_{1(t-1)} \\ \sigma_{11} \end{bmatrix} \sim NW^{-1}(m_{t-1}, c_{t-1} s_{11(t-1)}, d_{t-1}); \quad (6.15c)$$

Submodel 2:

$$y_{2t} = \theta_{2|1t} + y_{1t}\sigma_{12|1t} + e_{2|1t}, \quad e_{2|1t} \sim N(0, \sigma_{22|1}), \quad (6.16a)$$

$$\begin{bmatrix} \theta_{2|1} \\ \sigma_{12|1} \end{bmatrix}_t = \begin{bmatrix} g & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_{2|1} \\ \sigma_{12|1} \end{bmatrix}_{t-1} + \begin{bmatrix} f_{2|1} \\ 0 \end{bmatrix}_t, \quad \begin{bmatrix} f_{2|1} \\ 0 \end{bmatrix}_t \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} w & 0 \\ 0 & 0 \end{bmatrix} \sigma_{22|1}\right), \quad (6.16b)$$

$$\begin{bmatrix} \theta_{2|1t} \\ \sigma_{12|1t} \\ \sigma_{22|1} \end{bmatrix} \sim NW^{-1}\left(\begin{bmatrix} m_{2|1} \\ s_{12|1} \end{bmatrix}_{t-1}, \begin{bmatrix} c_{|1} & -m_{1}s_{11}^{-1} \\ -s_{11}^{-1}m_1 & s_{11}^{-1} \end{bmatrix}_{t-1}, s_{22|1(t-1)}, d_{t-1} + 1\right). \quad (6.16c)$$

Where the quantities in (6.16) are defined in accordance with the notation of Appendix A6.2, i.e. $\sigma_{12|1t} = \sigma_{12|1} = \frac{\sigma_{12}}{\sigma_{11}}, \theta_{2|1t} = \theta_{2t} - \theta_{1t}\sigma_{12|1}, \sigma_{22|1} = \sigma_{22} - \sigma_{21}\sigma_{12|1}, e_{2|1t} = e_{2t} - e_{1t}\sigma_{12|1}, f_{2|1t} = f_{2t} - f_{1t}\sigma_{12|1}$, etc.

It is clear from (6.15, 6.16) that the bivariate DWMR model (6.14) can be written as a dynamic linear recursive model. However, as a dynamic linear recursive model it has special characteristics. Firstly, the coefficient $\sigma_{12t} = \frac{\sigma_{12}}{\sigma_{11}}$ in (6.16) is static. Secondly, for implementation purposes, model (6.15) is redundant in the presence of (6.16), i.e. it is only necessary to implement the latter model. In particular the prior hyperparameters in (6.15c) are functions of the hyperparameters appearing in (6.16c). This relationship between the reduced form (6.14) and the recursive form (6.15, 6.16) suggests that the dynamic linear recursive model can be thought of as an extension of DLMR's with dynamic observational variance; see for example the form of Ω in (6.11). Conversely, a DLMR with dynamic scale variance, in a sense yet to be specified, could be a simple alternative to certain dynamic linear recursive models. In the rest of this section we explore these ideas in more detail.

6.3.1 Dynamic Stepwise Regression.

The DLMR model (3.5) can be rewritten, using the notation of Subsection 4.2.2 and Appendices A3.2, A6.2, in the following recursive form:

Submodel 1:

$$Y_{.1t} = X_t \Theta_{.1t} + E_{.1t}, \quad E_{.1t} \sim N(O, V_t, \Sigma_{11}), \quad (6.17a)$$

$$\Theta_{.1t} = G_t \Theta_{.2(t-1)} + F_{.1t}, \quad F_{.1t} \sim N(O, W_t, \Sigma_{11}), \quad (6.17b)$$

$$\begin{bmatrix} \Theta_{.1(t-1)} \\ \Sigma_{11} \end{bmatrix} \sim NW^{-1}(M_{.1(t-1)}, C_{t-1}, S_{11(t-1)}, d_{t-1}); \quad (6.17c)$$

Submodel 2:

$$Y_{.2t} = X_t \Theta_{.2|1t} + Y_{.1t} \Sigma_{12|1t} + E_{12|1t}, \quad E_{.2|1t} \sim N(O, V_t, \Sigma_{22|1}), \quad (6.18a)$$

$$\begin{bmatrix} \Theta_{.2|1} \\ \Sigma_{12|1} \end{bmatrix}_t = \begin{bmatrix} G_t & O \\ O & I \end{bmatrix} \begin{bmatrix} \Theta_{.2|1} \\ \Sigma_{12|1} \end{bmatrix}_{t-1} + \begin{bmatrix} F_{.2|1} \\ O \end{bmatrix}_t, \quad \begin{bmatrix} F_{.2|1} \\ O \end{bmatrix}_t \sim N \left(\begin{bmatrix} O \\ O \end{bmatrix}, \begin{bmatrix} W_t & O \\ O & O \end{bmatrix}, \Sigma_{22|1} \right), \quad (6.18b)$$

$$\begin{bmatrix} \Theta_{.2|1} \\ \Sigma_{12|1} \\ \Sigma_{22|1} \end{bmatrix}_{t-1} \sim NW^{-1} \left(\begin{bmatrix} M_{.2|1} \\ S_{12|1} \end{bmatrix}_{t-1}, \begin{bmatrix} C_{.1} & -M_{.1} S_{11}^{-1} \\ -S_{11}^{-1} M'_{.1} & S_{11}^{-1} \end{bmatrix}_{t-1}, S_{22|1(t-1)}, d_{t-1} + q_{.1} \right). \quad (6.18c)$$

Where

$$\Sigma_{12|1t} = \Sigma_{12|1} = \Sigma_{11}^{-1} \Sigma_{12}, \quad \Theta_{.2|1t} = \Theta_{.2t} - \Theta_{.1t} \Sigma_{12|1}, \quad \Sigma_{22|1} = \Sigma_{22} - \Sigma_{21} \Sigma_{12|1},$$

$$E_{.2|1} = E_{.2t} - E_{.1t} \Sigma_{12|1}, \quad F_{.2|1t} = F_{.2t} - F_{.1t} \Sigma_{12|1}, \quad \text{etc.}$$

The above result may be verified as follows. Equation (6.17a) and (6.18a) are simply another way of writing (3.5a), and their independence follows since

$$[E_{.1}, E_{.2|1}]_t = [E_{.1}, E_{.2}]_t \begin{bmatrix} I & -\Sigma_{12|1} \\ O & I \end{bmatrix} \sim N(O, V_t, \begin{bmatrix} \Sigma_{11} & O \\ O & \Sigma_{22|1} \end{bmatrix}).$$

Similarly, (6.17b) and (6.18b) are equivalent to (3.5b), and $[F_{.1}, F_{.2|1}]_t \sim N(O, W_t, \begin{bmatrix} \Sigma_{11} & O \\ O & \Sigma_{22|1} \end{bmatrix})$. Finally, result (A6.2.2b) implies that the prior (3.5c) can be decomposed into (6.17c) and (6.18c) and these priors are independent.

The implementation of the DLMR (3.5), according to (6.5) and (6.6) can be done via models (6.17) and (6.18). However, for this particular recursive model, the parallel implementation of (6.17) is redundant. Note the observational and evolution hyperparameters X_t, V_t, G_t and W_t are implicit in (6.18a) and (6.18b), and more importantly, the prior (and posterior) hyperparameters of model (6.17) always can be recovered from those of (6.18) via the sweep operation (A6.2.2c). The decomposition of (3.5) into (6.17) and (6.18) is valid for any partition of Y_t , thus the procedure for switching from one representation to another can be seen as a Bayesian (semi)dynamic counterpart of stepwise regression (see Appendix 4). Therefore, recalling that sweep operations are reflexive and order independent (see Appendix 4.1), (A6.2.2c) provides an effective procedure for including/excluding dependent variables as independent variables.

Applying decomposition (6.17, 6.18) repeatedly a fully recursive form of the DLMR is obtained. Thus, following the method described in Section 2, hierarchical missing data problems can be handled relatively easily. This method generalizes even in the static case, the procedure suggested by Chen (1986).

6.3.2 Discount Method.

A simple alternative already mentioned to certain dynamic recursive models is a DLMR with a dynamic scale variance. In the univariate case, the discount method provides a simple and natural way

for simulating a random walk type of evolution for σ^2 (Harrison and West, 1986). A straightforward multivariate generalization is characterized in terms of the distribution of Σ_t as follows.

Evolution:

$$\Sigma_t \sim W^{-1}(S_t^*, d_t^*). \quad (6.19)$$

where $S_t^* = \beta S_{t-1}$, $d_t^* = \beta d_{t-1}$, $\Sigma_{t-1} \sim W^{-1}(S_{t-1}, d_{t-1})$ and β is a discount factor representing a loss of $100(\beta^{-1} - 1)\%$ of the information about Σ_{t-1} in evolving to Σ_t . The distribution (6.19) represents our beliefs about Σ_t before observing Y_t . Thus, the updating recurrences induced by (6.19) is simply (3.10) after replacing S_{t-1}, d_{t-1} by S_t^*, d_t^* .

Doubts have been expressed (Dempster and Carlin, 1985) about the genuine Bayesian nature of discount methods in general. Nevertheless, discount methods produce a well defined joint distribution $p(Y_1) p(Y_2|Y_1) \dots p(Y_t|Y_1, \dots, Y_{t-1})$ for Y_1, Y_2, \dots, Y_t . Therefore, regarding the observables, discount methods are clearly coherent. Of course, completely arbitrary coherent procedures can also be defined; which brings us to the question of how can they be judged. A pragmatic answer is in terms of their predictive ability.

6.4 EXAMPLE: EXCHANGE RATE DYNAMICS.

We present here an application of the DWMR with a dynamic scale variance for modelling exchange rate dynamics. An abridged version may be found in Quintana and West (1986). This example illustrates the case, mentioned in Chapter 1, in which practical decision problems depend on the joint contemporaneous variation of the time series, and a main goal of the analysis is to learn about the unknown structure of such joint variation. This situation arises typically in relation to the spread of risk in investments, as in the design of optimal portfolios and motivates the analysis of the structure across several similar time series of price data; see, for example, Granger (1972).

6.4.1 Exchange rate Models.

The time series analysed are the exchange rates taken from the CSO macro-economic time series data bank. The time period studied is from January 1975 to August 1984 inclusive. Monthly exchange rates of the US dollar, the Deutschmark, the Japanese yen, the French franc, the Italian lira, and the Canadian dollar were considered, all relative to the British pound. This series plotted in log scale and standarized at the beginning of the time period for easy visual comparison, is shown in Figure 6.1. An important feature of this series is that the exchange rate policies, generally speaking, of the countries involved were flexible.

First, a DWMR of the form given by (3.31) is employed for modelling the series plotted in Figure 6.1, namely the shifted natural logarithms of the exchange rates. The logarithmic transformation is a conventional choice for analysing prices (Taylor, 1980; Fama, 1965). The cosmetic shift is irrelevant to the study because it is equivalent to a change of monetary units at the beginning of the time period.

Our basic DWMR model setting is:

$$\underline{y}_t = [y_{1t}, \dots, y_{6t}]', \quad \underline{x}_t = [1, 0]', \quad \Theta_t = \begin{bmatrix} \theta_{11} & \dots & \theta_{16} \\ \vdots & \ddots & \vdots \\ \theta_{21} & \dots & \theta_{26} \end{bmatrix}_t, \\ \Sigma = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{16} \\ \vdots & \ddots & \vdots \\ \sigma_{61} & \dots & \sigma_{66} \end{bmatrix}, \quad G_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Thus, each time series is marginally modelled by means of a dynamic linear trend with its own pair of time varying regression parameters. As usual, the contemporaneous variance covariance structure of the multivariate time series is essentially described by the system scale variance.

Several models can be accommodated in our basic DWMR depending on the setting of the driving parameters V_t, W_t (and the prior hyperparameters M_0, C_0, S_0, d_0). Two such cases follows:

(a) The naive, standard, multivariate regression with linear trend, reparameterized by means of a deterministic time variant parameter,

$$\underline{y}_t = \underline{\theta}_{1t} + \underline{\epsilon}_t, \quad \underline{\epsilon}_t \sim N(\underline{0}, \Sigma), \quad (6.20a)$$

$$\underline{\theta}_{1t} = \underline{\theta}_{1(t-1)} + \underline{\theta}_2, \quad (6.20b)$$

(with a vague prior) corresponds to the basic DWMR. The correspondence is given by $\Theta'_t = [\underline{\theta}_{1t}, \underline{\theta}_2]$, $v_t = 1, W_t = O$ (for all t) and $M_0 = 0, C_0 = I\epsilon^{-1}, S_0 = \epsilon I$ and $d_0 = \epsilon$ where $\epsilon \rightarrow 0^+$.

(b) The multivariate logarithmic discrete time version of the Brownian motion (Wiener process) model, proposed and studied in the pioneering work of Bachelier(1900),

$$\underline{y}_t = \underline{y}_{t-1} + \underline{f}_{1t}, \quad \underline{f}_{1t} \sim N(\underline{0}, \Sigma), \quad (6.21)$$

(with a vague prior) corresponds to the basic DWMR. The correspondence is given by $\Theta'_t = [\underline{\theta}_{1t}, \underline{0}]$ and $\underline{y}_t = \underline{\theta}_{1t}$ with the setting $v_t = 0, W_t = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, M_0 = O, C_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\epsilon^{-1}, S_0 = \epsilon I, d_0 = \epsilon$.

It is not difficult to see that generally the basic DWMR provides, depending on the choice of the hyperparameters, dynamic contenders for the trend random walk controversy; several points of view regarding this controversy may be found in Cootner (1964). The discrimination between (a finite number of) such rival models may be achieved by means of the multi-process models class I of Chapter 3. However, practical difficulties may arise depending on the dimension of \underline{y}_t (and of course, on the number of rival models). Some further discussion of this appears in later subsections.

6.4.2 Principal Components.

Principal component analysis of Σ requires, in principle, posterior distributions at each t for the eigenvalues/vectors of Σ . When Σ has an inverted Wishart distribution, these posteriors are extremely

difficult to work with. We restrict ourselves therefore, to consideration of point estimates of Σ over time. Proceeding as in Subsection 5.2.4 to estimate principal components, the value of the plug-in estimator of $\hat{\Sigma}$ given by (5.30) may be substituted into the defining equations,

$$\Sigma = P \Lambda P' \quad \text{and} \quad P' P = I, \quad (6.22)$$

in order to deduce estimates of the eigenvalues, forming the diagonal matrix Λ , and of the orthonormal eigenvectors, forming the columns of P . By convention we order the eigenvalues in decreasing order. For an initial illustration we report some features of the analysis in which $v_t = 1$, $S_0 = \epsilon I$, $d_0 = \epsilon = 10^{-5}$ and

$$W_t = \begin{bmatrix} .01 & 0 \\ 0 & .01 \end{bmatrix}, C_t = \begin{bmatrix} \epsilon^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

This, rather arbitrarily chosen, model allows for time variation in both level and growth parameters of the linear trend components, with magnitudes of such changes determined by W_t . Having processed all the 116 observations some features of the estimate $\hat{\Sigma}$ at time 116 are evident in the first three principal components shown in Table 6.2.

		(1)	(2)	(3)
U.S.A.	:	0.364	-0.555	0.010
Germany	:	0.398	0.210	-0.367
Japan	:	0.443	0.417	0.783
France	:	0.454	0.241	-0.256
Italy	:	0.387	0.189	-0.408
Canada	:	0.397	-0.622	0.094
% variation	:	65.6	14.6	12.0

Table 6.2 Weights of currencies in first 3 principal components

The first, dominant, component gives roughly equal weights to the six currencies, and thus represents an average performance index relative to the British pound; the resulting linear combination is a contender for a basket of currencies as a measure of Sterling value on international exchanges. In fact, after a renormalization this index virtually reproduces the exchange rate index as computed by the Bank of England. This is shown in Figure 6.3. The second component weights the E.E.C. countries roughly equally, adds in the Yen at about twice the weight, and contrasts this E.E.C./Japan aggregate with an average U.S.A./Canada index. The third component drops out the North American countries,

contrasting Japan with an average E.E.C. index. Time variation in the relative well-being of the three American/European/Japanese sectors be seen from the plot of the first three principal components, (using the final estimates in Table 1) over time. This is Figure 6.2.

6.4.3 Model Assessment.

Formal assessment of predictive performance of models using the cumulative product of observed one-step predictive densities can be applied to the entire multivariate model. For simplicity and economy, however, we assess models on the basis of forecasting performance on the univariate series obtained using the first principal component with weights as in Table 6.2. Since this component accounts for over 65% of total variation, predictive performance on this series should carry through to the original series. From (3.31) and (3.25) it is clear that any linear combination of the component series follows a simple univariate linear trend DLM. Within this single class of models, it is then easy to assess predictive performance using predictive densities.

Initially, the static linear trend (SLT) and random walk (RW) models, both as described in Subsection 6.4.1, may be compared using the Bayes' factor. The SLT is rapidly rejected in favor of RW which performs consistently better yielding a final cumulative Bayes' factor with a value of $\exp(189.2)$. A similar study of the entire multivariate series confirms this message coming from the first principal component. A further class of models of interest are those in which $v_t = v$,

$$W_t = \begin{bmatrix} w_{11} & 0 \\ 0 & w_{22} \end{bmatrix},$$

and $v + w_{11} + w_{22} = 1.02$ for all t . The latter constraint is for identifiability; a common factor in v and W is absorbed in Σ . Using similar, relatively vague priors, in such models for the first principal component series above, it turns out that predictive performance is optimized by RW-like models ($v = 0$), and in particular with models close to that with $v = 0$ and $w_{11} = 1, w_{22} = .02$. The one-step forecasts of this particular model appear in Figure 6.4. Its performance can be seen, in comparison with the initial model in example, in the plots of forecast residuals shown in Figure 6.5.

6.4.4 Dynamic Scale Variance.

There has been considerable discussion amongst financial time series modellers about the suggestion that series such as ours have variances that are essentially infinite. This has led, for example, to the use of stable distributions as alternatives to normality (Fama, 1985). It is absolutely clear, however, that such contentions depend entirely on the models used and within in which the variances have meaning. Use of an inflexible, static time series or regression models can easily lead to significant over-estimation of observational variances that may lead some investigators to suspect an infinite variance when, in fact, the fault lies elsewhere in the model. Harrison and West (1986) discuss such issues. Alternatively, if the structural form of the model is generally adequate, large variance estimates may derive from the assumption that observational variances are constant whereas, in fact, they are subject to change over time in a deterministic or stochastic manner. Such time dependent variances, and covariances

between series that change to reflect the changes in relationships over time, are not uncommon in economic and commercial applications (Harrison and West, 1986; Granger, 1972). In many cases, purely stochastic variation is evident and, often, a model allowing for slow, random changes in variances and covariances can adequately capture the important features of any structural changes and also allows for minor modelling errors. A simple and natural approach used here is the discount method discussed in Subsection 6.3.2.

The model chosen in Subsection 6.4.1 as optimal for the exchange rate series with constant Σ is re-examined and compared with an alternative in which the only difference is the use of a discount factor $\beta = .95$ rather than $\beta = 1$. The overwhelming weight of evidence in favor of the dynamic variance model is evident in the Bayes' factor for $\beta = .95$ versus $\beta = 1$, calculated overtime and plotted, on a logarithmic scale in Figure 6.6. Note that this is based on predictive densities for the full multivariate series rather than just the first principal component since that is no longer assumed stable over time in the dynamic model. The estimated $\hat{\Sigma}_t$ at time $t = 116$ from the dynamic scale variance model has the first three principal components given in Table 6.3.

		(1)	(2)	(3)
U.S.A.	:	0.228	-0.656	-0.122
Germany	:	0.457	0.269	-0.323
Japan	:	0.483	0.016	0.870
France	:	0.517	0.211	-0.273
Italy	:	0.433	0.135	-0.180
Canada	:	0.226	-0.658	-0.115
% variation	:	62.8	23.6	9.7

Table 6.3 Weights of currencies in first 3 principal components at time $t = 116$: dynamic scale variance model

The general features are similar to those in Subsection 6.3.2. There are, however, clear changes reflecting the need for a time varying Σ_t to adapt to changing economic relationships over the period of the data. Firstly, U.S.A. and Canada receive reduced weights in the basket of currencies provided by the first component, and in tandem the contrast between these two and the E.E.C. given in the second component has greater significance. Secondly, Japan is dropped from the second component, the relationship between Japan and North America coming in the third component where the latter are

now grouped with the E.E.C.

The time variant nature of Σ is noticeable in the sequential estimates at 6 months intervals, of the first three eigenvalues plotted in Figure 6.7.

6.4.5 Contemporaneous Conditional Form.

What-if questions such as: what would be the distribution of y_{2t} (Germany), y_{3t} (Japan) and y_{6t} (Canada) for a given y_{1t} (U.S.A.) at, say, $t = 117$? can be resolved directly by means of (4.15-4.16). Alternatively, a decomposition analogous to 6.17-6.18 can be employed. This latter approach not only provides a familiar parametric representation (and interpretation) but also allows for an analysis of the possible consequences before the actual y_{1t} is observed. Let us continue with the above example. We can identify Y_t in (3.5a) with $[y_{1t}, y_{2t}, y_{3t}, y_{6t}]$. Thus, the relevant parameters Θ_t, Σ_t , before observing y_t ($t = 117$), are distributed as

$$\begin{bmatrix} \Theta_t \\ \Sigma_t \end{bmatrix} \sim NW^{-1}(M_t^*, C_t^*, S_t^*, d_t^*),$$

where,

$$M_t^* = \begin{bmatrix} .599 & .390 & .808 & .325 \\ .0127 & .00253 & .00876 & .00908 \end{bmatrix}, \quad C_t^* = \begin{bmatrix} 1.15 & .152 \\ .152 & .172 \end{bmatrix},$$

$$S_t^* = \begin{bmatrix} .00723 & .00140 & .00283 & .0064 \\ .00140 & .00838 & .00568 & .00136 \\ .00283 & .00568 & .01109 & .00284 \\ .0064 & .00136 & .00284 & .00723 \end{bmatrix} \quad \text{and } d_t^* = 18.95.$$

In accordance with (A6.2.2), the parameters appearing in (6.18b), associated with $Y_{1,t} = y_{1t}$ and $Y_{2,t} = (y_{2t}, y_{3t}, y_{6t})$, are distributed as,

$$\begin{bmatrix} \Theta_{.2|1} \\ \Sigma_{12|1} \\ \Sigma_{22|1} \end{bmatrix}_t \sim NW^{-1} \left(\begin{bmatrix} M_{.2|1}^* \\ S_{12|1}^* \\ S_{22|1}^* \end{bmatrix}_t, \begin{bmatrix} C_{.1}^* & -M_{.1}^* S_{11}^{*-1} \\ -S_{11}^{*-1} M_{.1}^* & S_{11}^{*-1} \end{bmatrix}_t, S_{22|1t}^*, d_t^* + q_{.1} \right)$$

where

$$\begin{bmatrix} M_{.2|1}^* \\ S_{12|1}^* \\ S_{22|1}^* \end{bmatrix}_t = \begin{bmatrix} .273 & .573 & .205 \\ .0000587 & .00377 & -.0022 \\ .194 & .398 & .886 \end{bmatrix},$$

$$\begin{bmatrix} C_{.1}^* & -M_{.1}^* S_{11}^{*-1} \\ -S_{11}^{*-1} M_{.1}^* & S_{11}^{*-1} \end{bmatrix} = \begin{bmatrix} 50.9 & 1.21 & -82.9 \\ 1.21 & .194 & -1.76 \\ -82.9 & -1.76 & 138.4 \end{bmatrix},$$

$$S_{22|1t}^* = \begin{bmatrix} .00811 & .00513 & .000122 \\ .00513 & .00998 & .000336 \\ .000122 & .000336 & .00156 \end{bmatrix} \quad \text{and } d_t^* + q_{.1} = 19.95.$$

Inferences conditional on y_{1t} can be made now via (6.18a). This means that for a rise of one unit in the American index increments of .194, .398 and .886 are expected in the German, Japanese and Canadian indices respectively, etc.

Exchange Rates (Base: British Pound)

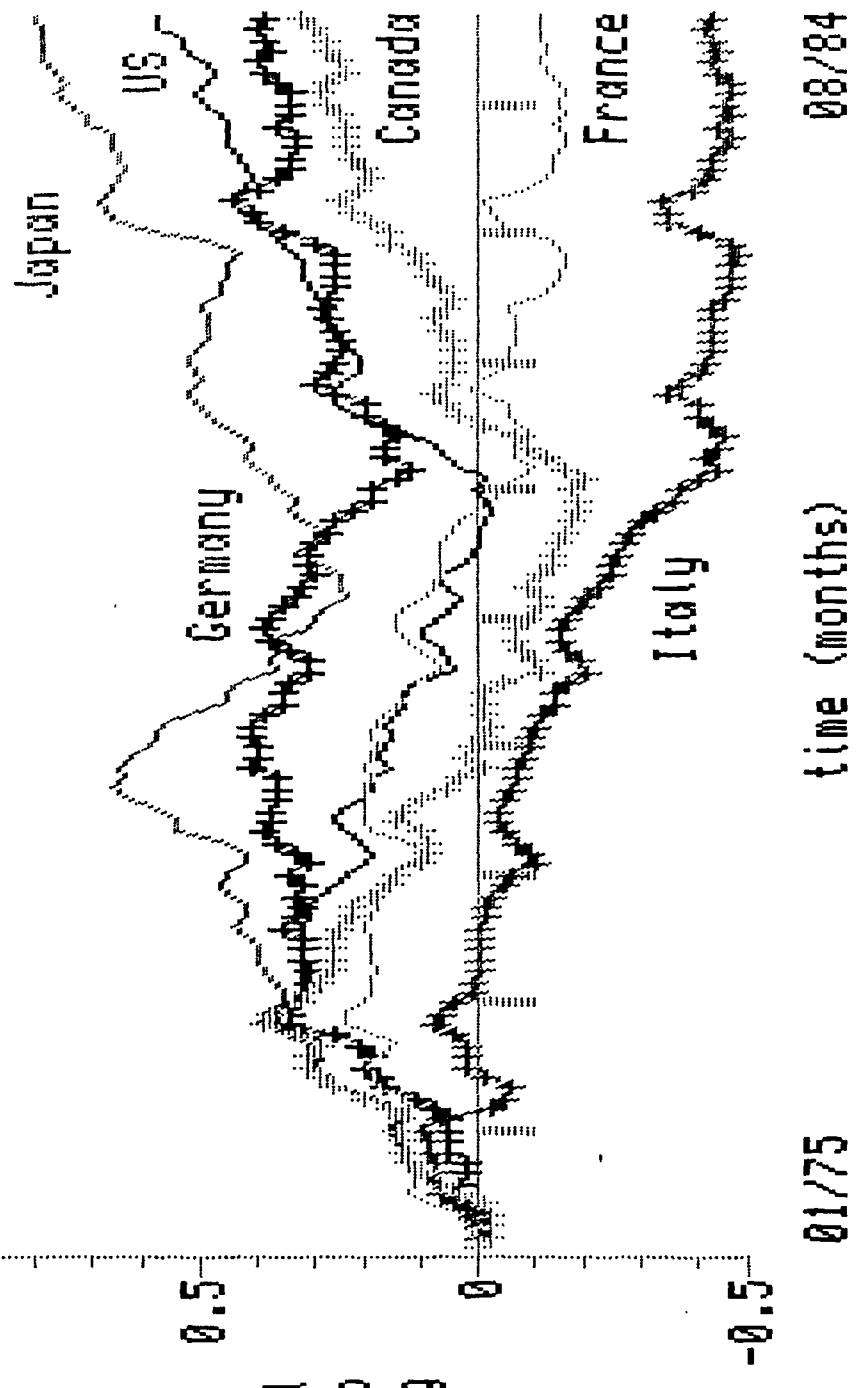


Figure 6.1 Exchange rates of the U.S. Dollar, the Deutschmark, the Japanese Yen, the French Frank, the Italian lira and the Canadian Dollar.

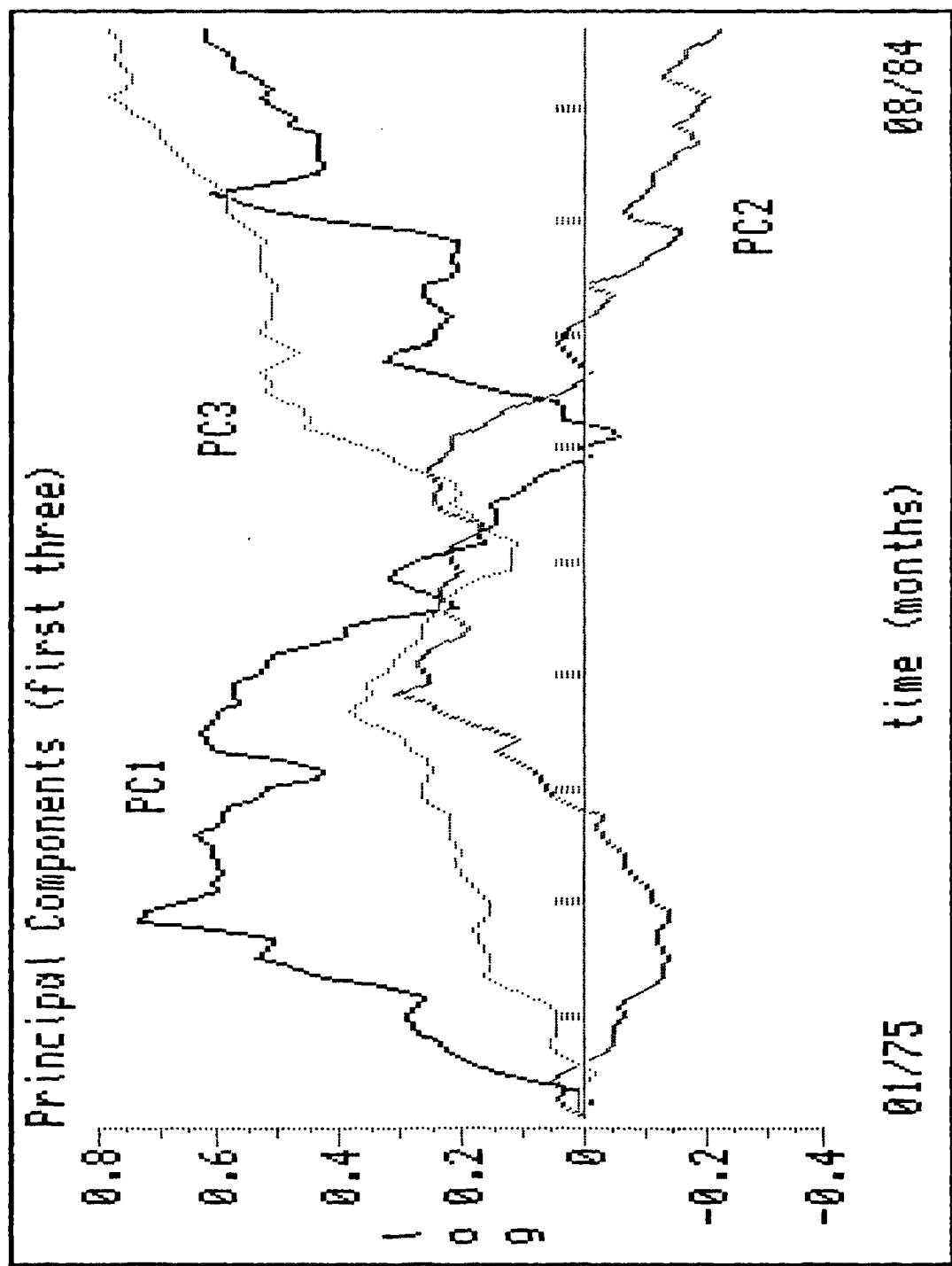


Figure 6.2 The first three principal components based on the initial model.

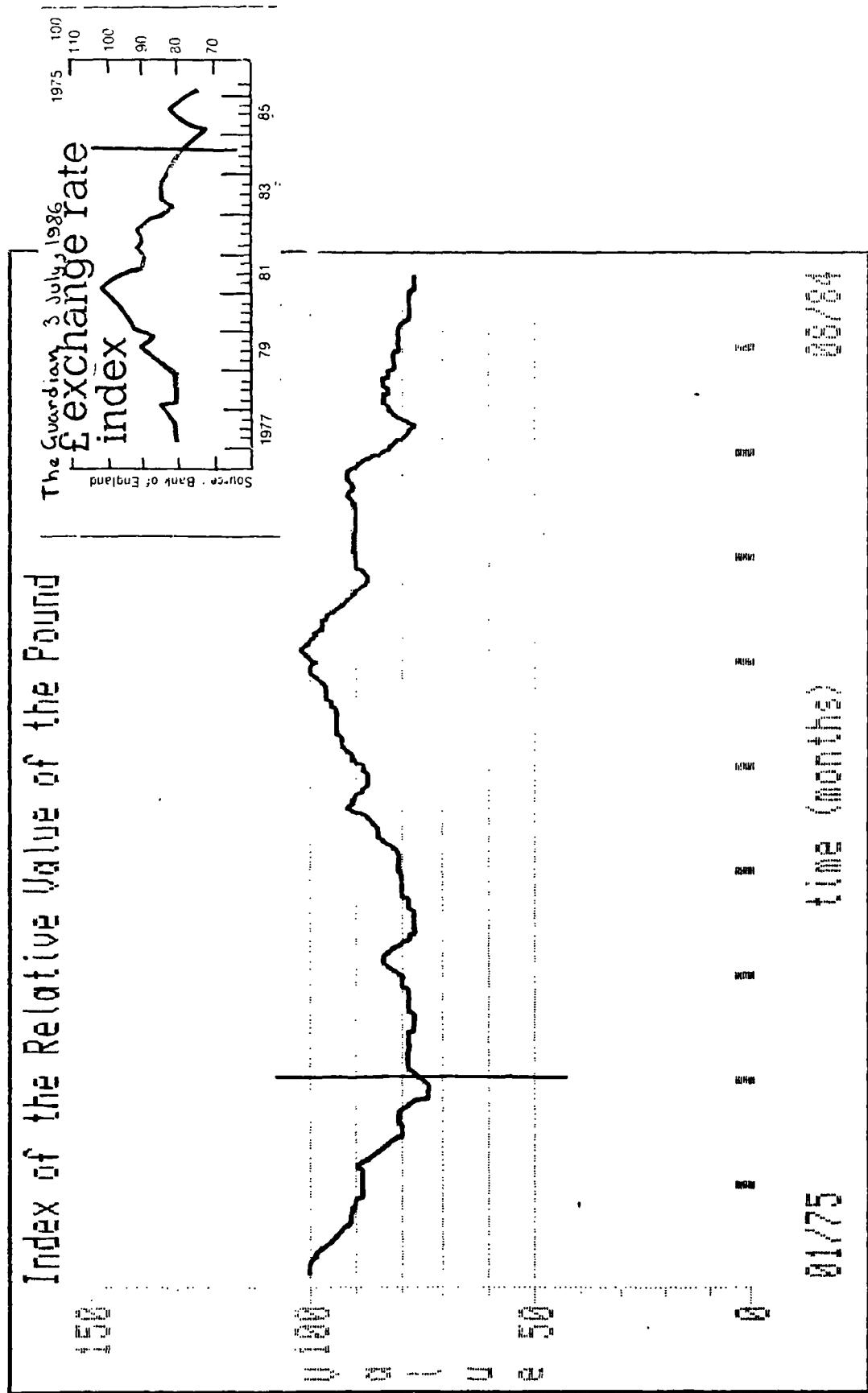


Figure 6.3 The (renormalized) first principal component as an index of the relative value of the British pound.

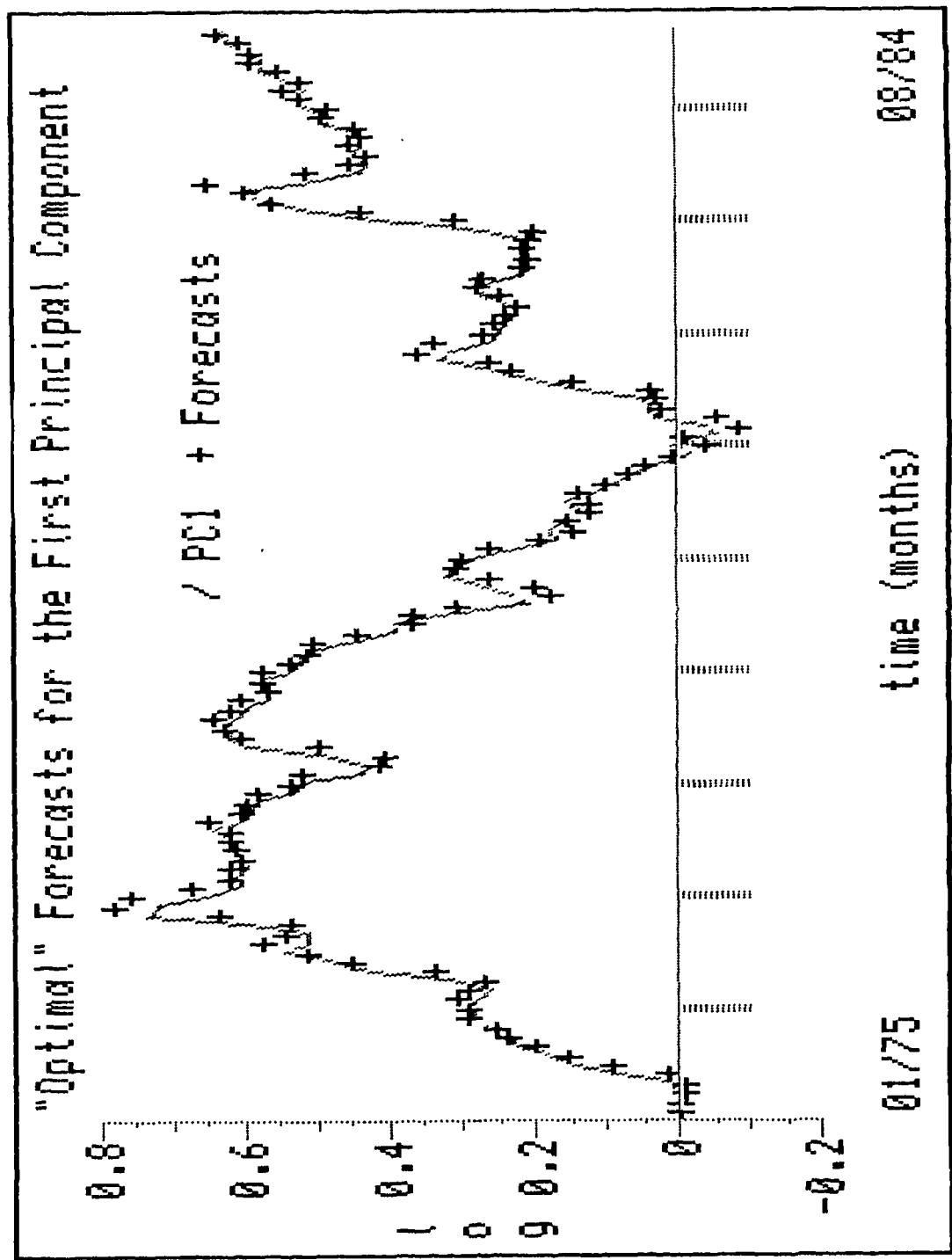


Figure 6.4 One-step forecasts for the first principal component using the 'optimal' model.

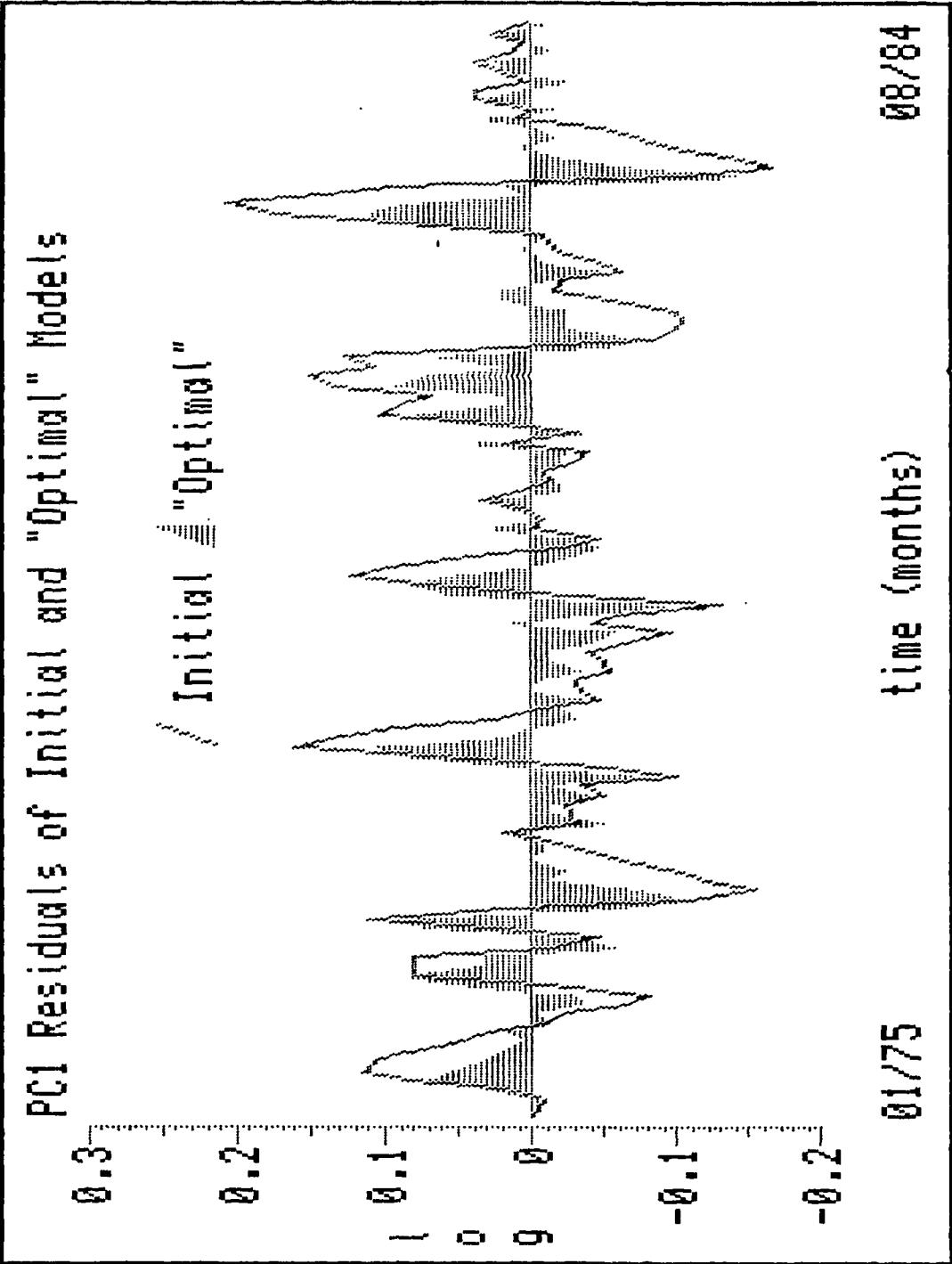


Figure 6.5 One-step residuals of the initial and 'optimal' models (for the first principal component).

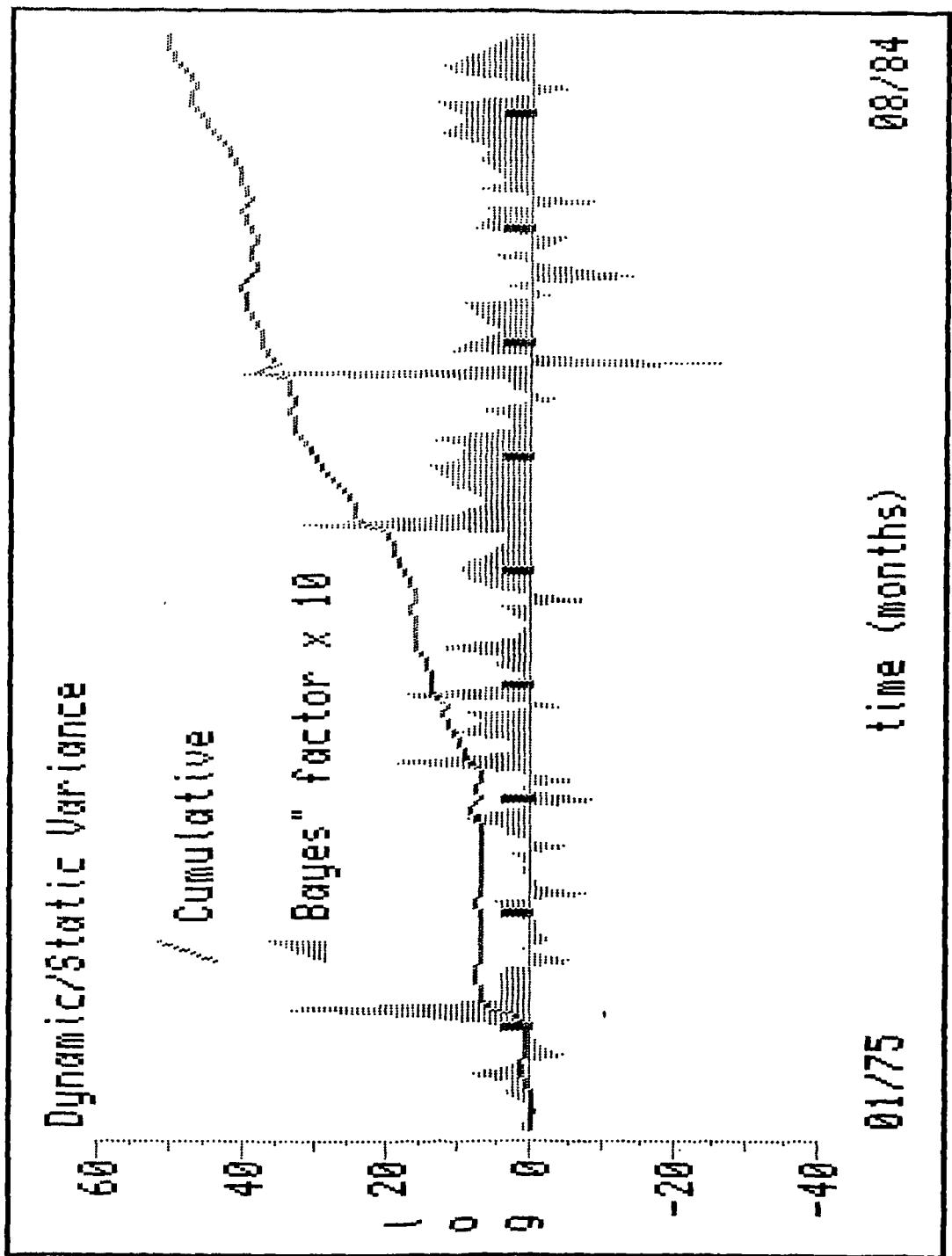


Figure 6.6 Bayes' factor and cumulative Bayes' factor for the 'optimal' model with dynamic variance relative to its static counterpart.

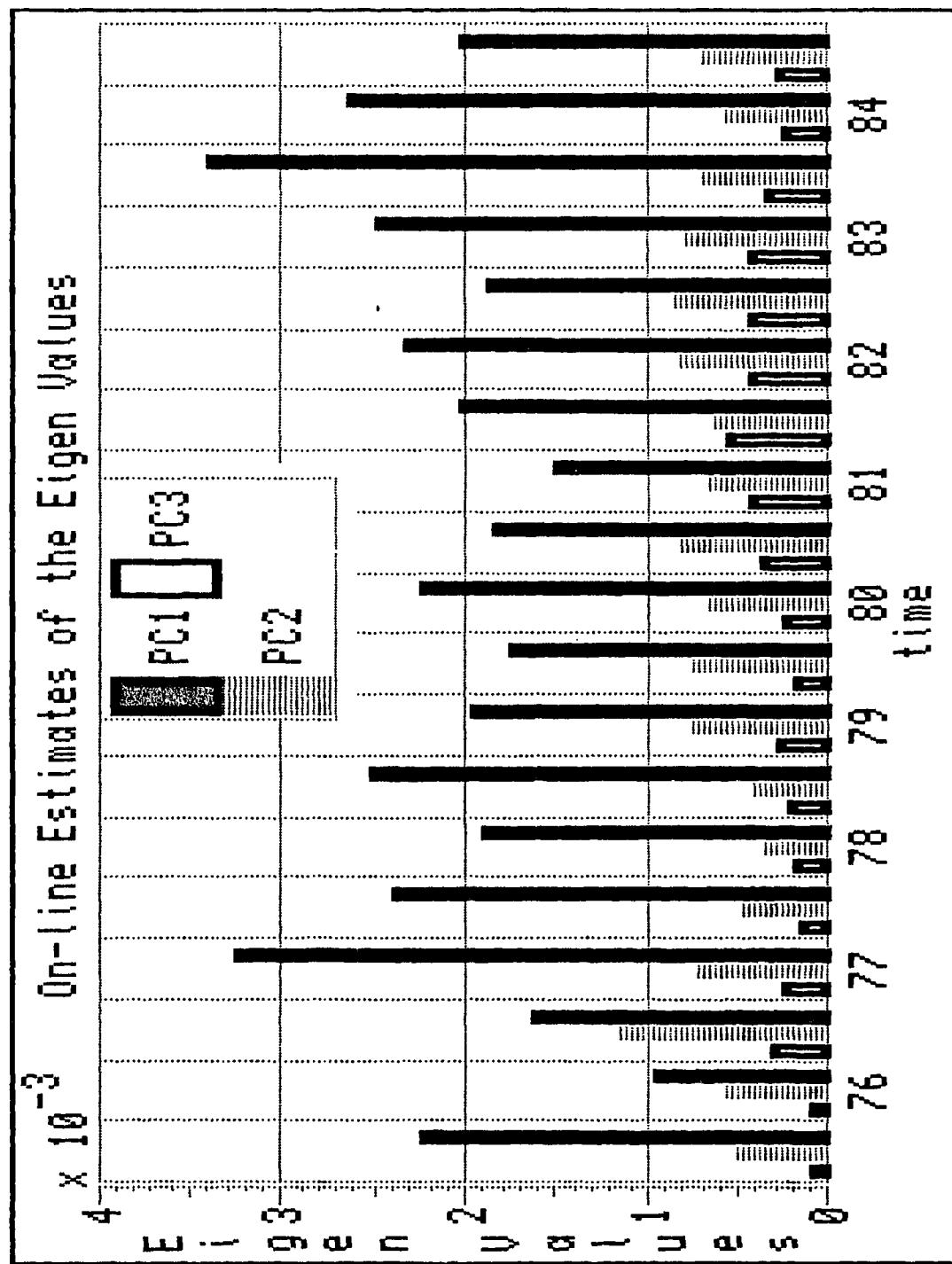


Figure 6.7 On-line estimates of the eigenvalues of Σ_t .

APPENDIX A6.1.

SIMULATION.

Bayesian analysis usually deals with random quantities of interest such as (functions of) unknown parameters and future observations for which there are simple procedures for generating random deviates, whereas their distributions are analytically intractable. The analysis of Section 2 is a typical example; all that we need in order to obtain a random sample of correlations are algorithms for generating deviates of normal and gamma random variables, and there are plenty of them. For generating normal deviates we may follow Marsaglia and Bray (1964):

- (1) Generate a pair of uniform deviates u_1 and u_2 .
- (2) Compute $y_1 = 2u_1 - 1$, $y_2 = 2u_2 - 1$, $s = y_1^2 + y_2^2$.
- (3) If $s \geq 1$ then reject the pair y_1, y_2 and goto step (1), otherwise compute $w = (-2 \frac{\log s}{s})$ and accept $x_1 = wy_1$ and $x_2 = wy_2$ as independent $N(0, 1)$ deviates.

Similarly, for generating gamma deviates with shape parameter $\alpha \geq 1$ (the usual case) we may follow Quintana (1985):

- (1) Generate a pair of uniform deviates u_1 and u_2 .
- (2) Compute $z = -\log u_1$.
- (3) If $(\alpha - 1)(1 - z + \log z) < \log u_2$ then reject z and goto step (1), otherwise accept $x = \alpha z$ as $\Gamma(\alpha, 1)$ deviate.

In fact, using these two algorithms it is possible to generate matrix-normal inverted-Wishart deviates. This may be shown (and implemented) by induction with the aid of result (A6.2.2b).

A common solution, when a sample generating algorithm is feasible, is to infer the properties of a distribution from the corresponding properties of a large, artificially generated random sample. In the rest of the appendix we discuss a Bayesian justification of this technique.

Let x be the random variable (scalar, vector or matrix) of interest and, as an illustration, let us suppose that we want to approximate the probability that x is in a subset A . The indicator variable $z = (x \in A)$ follows a Bernoulli(θ) where θ is the desired, unknown probability. Therefore, following the standard Bayesian conjugate analysis (De Groot, 1970, p. 160),

$$\theta \sim B(\alpha, \beta) \quad \text{implies } \theta|z \sim B(\alpha_z, \beta_z), \quad (\text{A6.1.1})$$

where B denotes the beta distribution, $z' = [z_1, \dots, z_n]$ is a random sample of size n from the Bernoulli, $\alpha_z = \alpha + n\bar{z}$, $\beta_z = \beta + n(1 - \bar{z})$, and \bar{z} denotes the sample mean $\frac{1}{n} \sum_{i=1}^n z_i$. Furthermore,

$$E_{\theta|z} \theta = \frac{\alpha_z}{\alpha_z + \beta_z} \quad \text{and} \quad \text{VAR}_{\theta|z} \theta = \frac{\alpha_z \beta_z}{(\alpha_z + \beta_z)^2 (\alpha_z + \beta_z + 1)}. \quad (\text{A6.1.2})$$

Since we can be sampling virtually as much as we want the likelihood will dominate the prior. When using a vague prior $\alpha \rightarrow 0, \beta \rightarrow 0$ equations (A6.1.2) result in,

$$E_{\theta|z} \theta = \bar{z} \quad \text{and} \quad \text{VAR}_{\theta|z} \theta = \frac{\bar{z}(1 - \bar{z})}{n + 1} \leq \frac{1}{4(n + 1)}. \quad (\text{A6.1.3})$$

Thus, \bar{z} is an estimate of θ and we can choose in advance the sample size in order to assure a desired precision.

The partition of the sample space in the above discussion is very simple: $\{A, A^c\}$. When the probabilities of interest correspond to a finite partition $\{A_i\} (i = 1, \dots, n)$, the above Bernoulli-beta model may be extended to the standard multinomial-Dirichlet (De Groot, 1970, p. 174). Here the desired, unknown probabilities are θ_i with associated prior Dirichlet parameters α_i (i.e. $p(\theta_1, \dots, \theta_n) \propto \prod_{i=1}^n \theta_i^{\alpha_i - 1}$) and indicator variables $z_i = (x \in A_i)$. Several partitions can be embedded consistently by collapsing the corresponding components. For instance, consider two partitions $\{A_i\} (i = 1, \dots, n)$, $\{B_j\} (j = 1, \dots, m)$ and the refinement $\{C_{ij}\}$ where $C_{ij} = A_i B_j$. Then a multinomial-Dirichlet model defined on $\{C_{ij}\} (\theta_{ij}, \alpha_{ij}, z_{ij})$ induces multinomial-Dirichlet models on $\{A_i\} (\theta_i, \alpha_i, z_i)$ and $\{B_j\} (\theta_j, \alpha_j, z_j)$, where $\theta_i = \sum_{j=1}^m \theta_{ij}$, $\theta_j = \sum_{i=1}^n \theta_{ij}$, $\alpha_i = \sum_{j=1}^m \alpha_{ij}$, $\alpha_j = \sum_{i=1}^n \alpha_{ij}$, and $z_i = \sum_{j=1}^m z_{ij}$, $z_j = \sum_{i=1}^n z_{ij}$. Furthermore, multinomial-Dirichlet models corresponding to any finite partitions can be embedded consistently in a Dirichlet process model. In addition, other quantities of interest can be approximated, for instance, it can be shown that, assuming vague prior information, the best estimator with square error loss of the mean of a function of x is the corresponding sample mean, the details can be found in Ferguson (1973).

Monte Carlo integration and simulation techniques in particular have been criticized in O'Hagan (1986). The main objection is that Monte Carlo ignores information contained in the sample itself and therefore it wastes available information. However, for large samples this objection seems rather weak.

APPENDIX A6.2.

DISTRIBUTION OF SWEEP MATRICES.

We present here two results concerning the distribution of swept matrices which provide the basis for dynamic stepwise regression. We follow, naturally, the notation of Appendices A3.2 and A4.1.

The first result characterizes the inverted-Wishart distribution by way of a sweeping operation,

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \longleftrightarrow \begin{bmatrix} \Sigma_{11}^{-1} & \Sigma_{12|1} \\ -\Sigma'_{12|1} & \Sigma_{22|1} \end{bmatrix}, \quad (\text{A6.2.1a})$$

stating that,

$$\begin{aligned} \Sigma &\sim W^{-1}(S, d) \quad \text{if and only if} \\ \Sigma_{11}^{-1} &\leftrightarrow \Sigma_{11} \sim W^{-1}(S_{11}^{-1}, d) \quad \text{and, independently,} \\ \begin{bmatrix} \Sigma_{12|1} \\ \Sigma_{22|1} \end{bmatrix} &\sim NW^{-1}(S_{12|1}, S_{11}^{-1}, S_{22|1}, d + q_1). \end{aligned} \quad (\text{A6.2.1b})$$

The parameters are defined by the counterpart of (A6.2.1a),

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \longleftrightarrow \begin{bmatrix} S_{11}^{-1} & S_{12|1} \\ -S'_{12|1} & S_{22|1} \end{bmatrix}, \quad (\text{A6.2.1c})$$

and q_1 is the order of the pivoting matrix Σ_{11} . The derivation of this result may be found in Dempster (1969, p. 297-298) and in Box and Tiao (1973, p. 461).

The second result generalizes the first to the matrix-normal inverted-Wishart distribution, which is characterized via,

$$\begin{bmatrix} \Theta_{.1} & \Theta_{.2} \\ \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \longleftrightarrow \begin{bmatrix} -\Theta_{.1}\Sigma_{11}^{-1} & \Theta_{.2|1} \\ \Sigma_{11}^{-1} & \Sigma_{12|1} \\ -\Sigma'_{12|1} & \Sigma_{22|1} \end{bmatrix}. \quad (\text{A6.2.2a})$$

We have

$$\begin{aligned} \begin{bmatrix} \Theta \\ \Sigma \end{bmatrix} &\sim NW^{-1}(M, C, S, d) \quad \text{if and only if} \\ \begin{bmatrix} -\Theta_{.1}\Sigma_{11}^{-1} \\ \Sigma_{11}^{-1} \end{bmatrix} &\leftrightarrow \begin{bmatrix} \Theta_{.1} \\ \Sigma_{11} \end{bmatrix} \sim NW^{-1}(M_{.1}, C, S_{11}, d) \quad \text{and, independently,} \\ \begin{bmatrix} \Theta_{.2|1} \\ \Sigma_{12|1} \\ \Sigma_{22|1} \end{bmatrix} &\sim NW^{-1} \left(\begin{bmatrix} M_{.2|1} \\ S_{12|1} \end{bmatrix}, \begin{bmatrix} C_{|.1} & -M_{.1}S_{11}^{-1} \\ -S_{11}^{-1}M'_{.1} & S_{11}^{-1} \end{bmatrix}, S_{22|1}, d + q_1 \right) \end{aligned} \quad (\text{A6.2.2b})$$

The parameters are defined by the counterpart of (A6.2.2a),

$$\begin{bmatrix} C & M_{.1} & M_{.2} \\ -M'_1 & S_{11} & S_{12} \\ -M'_2 & S_{21} & S_{22} \end{bmatrix} \leftrightarrow \begin{bmatrix} C_{|.1} & -M_{.1}S_{11}^{-1} & M_{.2|1} \\ -S_{11}^{-1}M'_{.1} & S_{11}^{-1} & S_{12|1} \\ -M'_{.2|1} & -S'_{12|1} & S_{22|1} \end{bmatrix}. \quad (\text{A6.2.2c})$$

Identifying $\begin{bmatrix} \Theta \\ \Sigma \end{bmatrix}$ in (A6.2.2b) with $\begin{bmatrix} \Sigma_{12|1} \\ \Sigma_{22|1} \end{bmatrix}$ in (A6.2.1a) it is clear that, for d (in A6.2.2b) greater than the order of C , result (A6.2.1b) implies (A6.2.2b) since, given M, C and S ,

$$\begin{bmatrix} C & M \\ -M' & S \end{bmatrix} \leftrightarrow \begin{bmatrix} C^{-1} & C^{-1}M \\ M'C^{-1} & S + M'C^{-1}M \end{bmatrix}, \quad (\text{A6.2.3})$$

and the matrix in the right-hand side of (A6.2.3) is positive definite. Furthermore (A6.2.2b) can be derived directly (for any $d > 0$), following an analogous pattern of the proof of (A6.2.1b) found in Box and Tiao (1973, p. 462). The derivation is not difficult but it is rather cumbersome and it does not provide any further insight, hence it is omitted.

It is important to note that we have assumed implicitly that the distributions are non-singular. However, we can overcome this problem proceeding sequentially as usual.

CHAPTER 7

MODELLING ASPECTS

This chapter concerns various modelling aspects of the DWMR and some special models contained within the general case. The setting of vague priors is discussed in Section 6.1. The use of two very important multivariate transformations, logarithmic and logarithmic ratio, is examined in Section 6.2. Finally, Section 6.3 deals with alternative but equivalent reformulations of the DLMR: perfect observations, colored observation errors, colored evolution noises, correlated noise and error, fixed-lag smoothing and prediction, differencing series and transfer response functions.

7.1 VAGUE PRIORS.

A characteristic of the Bayesian approach is that the prior provides a means of incorporating valuable information. The setting of the initial distribution of the parameters in a DLMR is far from trivial; however, the methods of Dickey, Dawid and Kadane (1986) may be helpful. Nevertheless, there are situations in which the information about the parameters at time $t = 0$ is vague, or we want, for convenience, to perform the analysis as if this were the case. Equations (3.7c) and (3.9) represent the changes in the hyperparameters as the information about Θ_t and Σ increases by observing Y_t , and they suggest the following limiting values for the inverse process: $C_0^{-1} \rightarrow O$, $S_0 \rightarrow O$, $d_0 \rightarrow 0$ and, say, $M_0 = O$. This effect can be achieved by setting $M_0 = O$, $C_0 = \epsilon^{-1}I$, $S_0 = \epsilon I$ and $d_0 = \epsilon \rightarrow 0$. Notice that, in practice, the recurrences (3.7c) and (3.9) or alternatively the pivoting transformation (4.10) still can be performed by setting ϵ equal to a small positive value, say the square root of the particular machine precision.

Vague priors of the type mentioned above provide a Bayesian interpretation of several non-Bayesian results. This is discussed in Subsection 4.1.4. Other starting values for d_0 are possible, for instance, a procedure based on Jeffreys' rule yields $d_0 = -q + 1$; see Box and Tiao (1973, p. 426). Related to the use of vague priors is the problem of overfitting. Typically, a vague prior produces null values associated with S_t for $t = 0, 1, \dots, p$ and creates an illusion of accurate forecasts when, in fact, it is the effect of the prior. In consequence, great care is necessary when comparing models with a different number of regression parameters (p). The previous discussion suggests yet another possible starting value for the shape parameter: $d_0 = -p$.

7.2 TRANSFORMATIONS.

It is well known that the scope of applications for linear normal modes is significantly widened by means of transforming the data. In this section we briefly discuss two very important multivariate transformations in the context of DWMR models.

7.2.1 Log Transformation.

This transformation consists of taking the natural (by convention) logarithms of the original series. This procedure transforms a multivariate time series defined over a high-dimensional positive orthant into a new time series defined over the corresponding high-dimensional real space, this resulting multivariate time series is then modelled via a DWMR. Thus, the observational distribution of the original series is implicitly assumed to be a multivariate log-normal distribution (see Press, 1982, p. 148-150). Moreover, it is clear from (3.25) that any geometric combination of the original series also belongs to the same class of models and, in addition, the corresponding parameters X_t, V_t, G_t and W_t are invariant.

As a general guide-line we can say that a log transformation is recommended when the series' behaviour is better explained in terms of proportions rather than in terms of increments. A typical example is found in Section 5.2.

7.2.2 Log Ratio Transformation.

The log ratio transformation is intrinsically related to the logarithmic transformation, and consists of taking natural logarithms of the ratios between each univariate series and a fixed reference series, i.e. assuming that $\underline{z}_t = (z_{1t}, \dots, z_{qt})$ is the original series and, say z_{qt} is the reference series, then the transformed series $\underline{y}'_t = (y_{1t}, \dots, y_{qt})$ is given by,

$$y_{jt} = \log \frac{z_{jt}}{z_{qt}} = \log z_{jt} - \log z_{qt}, \quad j = 1, 2, \dots, q. \quad (7.1)$$

This procedure transforms a multivariate time series defined over the $(q-1)$ -dimensional positive simplex into a new time series defined over the $(q-1)$ -dimensional real space, again, the resulting series is then modelled via a DWMR. The inverse of the log ratio transformation (7.1) is the logistic transformation,

$$z_{jt} = \frac{\exp(y_{jt})}{\sum_{j=1}^q \exp(y_{jt})}, \quad j = 1, \dots, q. \quad (7.2)$$

This implies that the observational distribution of the original series is a multivariate logistic-normal distribution. Many of the well-known properties of the logistic-normal distribution, as set out for example in Aitchison and Shen (1980) may be extended to the dynamic case. The counterpart of (3.25) is that any standarized geometrical combination of ratios of the original series also belongs to the same class of models and the parameters X_t, V_t, G_t and W_t are invariant. This shows, in particular, that the analysis is invariant with respect to the order in which the series components are considered and therefore the choice of the reference series is irrelevant. Furthermore, switching from one representation to another can be easily accomplished by means of (3.25). For instance, let \underline{w}'_t be the transformed time series taking z_{1t} as the reference series, then \underline{w}'_t and \underline{y}'_t are related by,

$$\underline{w}'_t = \underline{y}'_t \begin{bmatrix} 0 & -1' \\ 0 & I \end{bmatrix}. \quad (7.3)$$

An equivalent symmetric transformation (Aitchison, 1983) uses the geometrical mean as a reference series instead of a particular component. In so doing, the asymmetric constraint in which a particular

component of the resulting series is identically null is replaced by,

$$\underline{s}'_t \underline{1} = 0, \quad (7.4)$$

where \underline{s}'_t is the transformed series. This symmetric representation may be obtained from any asymmetric representation as follows,

$$\underline{s}'_t = \underline{w}'_t K = \cdots = \underline{y}'_t K, \quad (7.5)$$

where $K = I - \underline{1} q^{-1} \underline{1}'$. Again, notice that X_t, V_t, G_t and W_t are invariant.

It is interesting to note that a log ratio DWMR is, in fact, employed in the example of Section 6.4. Admittedly, the corresponding proportions are forced, rather arbitrarily, to be even at the beginning of the time period; however, the obvious focus of attention, in that example, is not on the proportions themselves, but on how they change over time. Moreover, a formula like (7.3) together with (3.25) may be employed for switching from the model based on the U.K. pound to another based on, say, the U.S. dollar.

The relationship between the log DWMR and the log ratio DWMR is twofold. First, the relative ratios of a series modelled via a log ratio DWMR constitute a series implicitly modelled via a log DWMR. Conversely, the proportions of a series modelled via a log DWMR form a series implicitly modelled via a log ratio DWMR.

3. SPECIAL MODELS.

Useful dynamic models with particular characteristics represent certain time series more adequately than the DLMR specified by (3.5); however, these alternative models are, in fact, equivalent to the DLMR's, i.e. after some manipulation they can be reformulated as DLMR's. Equipped with this rich stock of models, including mixtures, the field of applications is 'bown wide open. These models with well-known counterparts in DLM's and/or state-space formulations are outlined below.

(a) Perfect Observations.

In this model there is no observation error, so that its observation equation is reduced to,

$$Y_t = X_t \Theta_t. \quad (7.6)$$

Clearly this model is equivalent to a DLMR with $V_t = O$. An interesting characteristic of perfect observation models is that the posterior distributions are always singular. One can come to the conclusion that since this model is a naive particular case of the DLMR, it should be of little interest; however, this impression may be misleading as is shown below.

(b) Colored Observation Error.

In this model the assumption of independence of E_t over time is replaced by a Markovian equation

$$E_t = D_t E_{t-1} + H_t, \quad H_t \sim N(O, V_t, \Sigma), \quad (7.7)$$

where D_t is known. For $D_t = O(t = 1, 2, \dots)$ this model becomes the usual DLMR (3.5). On the other hand, it can be reformulated as a DLMR with perfect observations by redefining the observation and evolution equations as,

$$Y_t = [X, I]_t \begin{bmatrix} \Theta \\ E \end{bmatrix}_t, \quad (7.8a)$$

$$\begin{bmatrix} \Theta \\ E \end{bmatrix}_t = \begin{bmatrix} G & O \\ O & D \end{bmatrix}_t \begin{bmatrix} \Theta \\ E \end{bmatrix}_{t-1} + \begin{bmatrix} F \\ H \end{bmatrix}_t, \quad (7.8b)$$

where $\begin{bmatrix} F \\ H \end{bmatrix}_t \sim N\left(\begin{bmatrix} O \\ O \end{bmatrix}, \begin{bmatrix} W & O \\ O & V \end{bmatrix}_t, \Sigma\right)$.

(c) Colored Evolution Noise.

This model is the evolutional counterpart of (b). The Markovian evolution for F_t is,

$$F_t = A_t F_{t-1} + B_t, \quad B_t \sim N(O, W_t, \Sigma), \quad (7.9)$$

where A_t is known. The equivalence with the DLMR can be shown by considering the following observation and evolution equations,

$$Y_t = [X_t, O] \begin{bmatrix} \Theta_t \\ F_{t+1} \end{bmatrix} + E_t, \quad (7.10a)$$

$$\begin{bmatrix} \Theta_t \\ F_{t+1} \end{bmatrix} = \begin{bmatrix} G_t & I \\ O & A_{t+1} \end{bmatrix} \begin{bmatrix} \Theta_{t-1} \\ F_t \end{bmatrix} + \begin{bmatrix} O \\ B_{t+1} \end{bmatrix}. \quad (7.10b)$$

(d) Correlated Noise and Error.

In this model, as its name implies, the assumption of independence of E_t and F_t is replaced by

$$\begin{bmatrix} F \\ E \end{bmatrix}_t \sim N\left(\begin{bmatrix} O \\ O \end{bmatrix}, \begin{bmatrix} W & U \\ U' & V \end{bmatrix}_t, \Sigma\right), \quad (7.11)$$

where U is known. This model is essentially (b) with an obvious modification.

(e) Fixed-lag Smoothing and Prediction.

Smoothing and prediction concern the problem of obtaining the distribution of the regression parameters conditional on the past and present information. Depending on whether the time interest is past or future the problem is referred to as smoothing or prediction (Gelb, 1974; Maybeck 1982a). Fixed-lag smoothing and prediction can be achieved easily by augmenting the regression parameters in order to include the desired past and/or future regression parameters. For example, suppose that the one-step ahead and back regression parameters are of interest. The observation and evolution equations are,

$$Y_t = [O, X_t, O] \begin{bmatrix} \Theta_{t-1} \\ \Theta_t \\ \Theta_{t+1} \end{bmatrix} + E_t, \quad (7.12a)$$

$$\begin{bmatrix} \Theta_{t-1} \\ \Theta_t \\ \Theta_{t+1} \end{bmatrix} = \begin{bmatrix} O & I & O \\ O & O & I \\ O & O & G_{t+1} \end{bmatrix} \begin{bmatrix} \Theta_{t-2} \\ \Theta_{t-1} \\ \Theta_t \end{bmatrix} + \begin{bmatrix} O \\ O \\ F_{t+1} \end{bmatrix}. \quad (7.12b)$$

Fixed-lag observational forecasts may be obtained similarly by regarding the future observations as observable regression parameters.

(f) Differencing Series.

It is a wide spread practice in time series analysis to overcome some difficulties associated with ill-behaved series by differencing the data. This differencing process can be achieved within the model itself by considering the DLMR with perfect observations described by the following observation and evolution equations,

$$Y_t = [I, O] \begin{bmatrix} Y_t \\ \Theta_{t+1} \end{bmatrix}, \quad (7.13a)$$

$$\begin{bmatrix} Y_t \\ \Theta_{t+1} \end{bmatrix} = \begin{bmatrix} I & X_t \\ O & G_{t+1} \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ \Theta_t \end{bmatrix} + \begin{bmatrix} E_t \\ F_{t+1} \end{bmatrix}. \quad (7.13b)$$

See, for example, the RW-like models of Subsection 6.4.3. The procedure for higher order differencing is easily appreciated.

(g) Transfer Response.

Transfer functions model the response to independent variables Z_t such as prices, advertising, etc. This effect is incorporated into a full model via the superposition principle. Transfer response modelling in the Bayesian forecasting context is discussed in Migon (1984, Chapter 6). A transfer response component for a DLMR is defined by the observation and evolution equations,

$$Y_t = [X_t, O] \begin{bmatrix} \Theta_t \\ \Psi_{t+1} \end{bmatrix}, \quad (7.14a)$$

$$\begin{bmatrix} \Theta_t \\ \Psi_{t+1} \end{bmatrix} = \begin{bmatrix} G_t & Z_t \\ O & I \end{bmatrix} \begin{bmatrix} \Theta_{t-1} \\ \Psi_t \end{bmatrix} + \begin{bmatrix} O \\ K_{t+1} \end{bmatrix}, \quad (7.14b)$$

where Ψ_t is the time varying gain and K_t is its noise. For instance, a trivariate DWMR transfer response component with an exponential decay factor λ is represented by,

$$[y_1, y_2, y_3]_t = [\theta_1, \theta_2, \theta_3]_t,$$

$$\begin{bmatrix} \theta_{1t} & \theta_{2t} & \theta_{3t} \\ \psi_{1(t+1)} & \psi_{2(t+1)} & \psi_{3(t+1)} \end{bmatrix} = \begin{bmatrix} \lambda & z_t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_{1(t-1)} & \theta_{2(t-1)} & \theta_{3(t-1)} \\ \psi_{1t} & \psi_{2t} & \psi_{3t} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ k_{1(t+1)} & k_{2(t+1)} & k_{3(t+1)} \end{bmatrix},$$

where $0 < \lambda < 1$.

CHAPTER 8

DISCUSSION AND FURTHER RESEARCH

In the preceding chapters the Bayesian forecasting approach has been employed as a means of introducing and developing flexible and tractable multivariate time series models. In so doing, we have attempted to apply related techniques in an innovative way for solving practical problems. As a close to the thesis we now discuss our results, and point out topics where further research is required.

The DLMR model is a general, easy to implement, matrix-variate DLM. The DLMR not only provides an on-line learning procedure for the regression parameters but also for the scale variance matrix. Moreover, it is a very flexible tool since it inherits the facilities of the conventional DLM: the use of prior information, construction of complex models via the superposition principle, intervention analysis, etc.

The row vector DLMR, i.e. the DWMR, is very useful for modelling multivariate time series when the components can be described by univariate DLM's with a similar structure. In such case, the scale matrix variance is essentially the observational variance, and the model offers a powerful procedure for learning about it. Nevertheless, the DWMR cannot cope when the component series have different structural behaviours, common repression parameters, or when their regression parameters interact with each other in the evolution equations. All these cases, however, can be handled by the column vector DLMR, i.e. the multivariate DLM. The price to be paid for this generality is that the learning procedure for the variance-covariance structure is very limited because the scale variance matrix in this case is a simple scalar. The strict matrix-variate DLMR is a compromise between the DWMR and the multivariate DLM. In this regard, it would be interesting to look at an application. For instance, if the multivariate DLM is able to model certain economic time series of one country, then a DLMR may be employed for modelling simultaneously the same kind of time series for several countries with similar economic systems.

We recommend highly the implementation of the DLMR updating recurrences via the sweep operator. The DLMR can cope with not only singular and non-singular models but also can be programmed easily. The generalized state-space filters of Chapters 4 can be modified in order to deal with non-singular models as well. However, it is necessary to compute pseudo-inverses of matrices, and therefore a special algorithm is required. In fact, the sweep operator can be used for that purpose (Goodnight, 1979), but to do so would be similar to employing Gaussian elimination for computing inverses in order to solve a system of simultaneous equations. Therefore, the direct application of the sweep operator as shown in Chapter 4 is preferred. In addition, the sweep operator has other theoretical and practical applications in the DLMR context, for example in dynamic step-wise regression and in contemporaneous conditional predictive distributions.

Multi-process modelling is a versatile tool passed on to the DLMR from the DLM. Multi-process modelling class I is useful for trimming the hyperparameters through the consideration of a fixed collection of models. For instance, it can be employed for dealing with non-standard priors. Virtually any prior distribution of the parameters of the DLMR can be approximated, in principle, by a mixture (linear convex combination) of matrix-normal inverted-Wishart distributions. In practice, we need to consider only the prior distributions that can be adequately represented by a mixture with a reasonably small number of components. When external information is critical, as in intervention analysis, a substantial forecasting improvement may be achieved by enriching the prior distributions in this way. Multi-process modelling class II handles the situation in which the possible models may change from one interval to another. However, the implementation of these multi-process models is, in general, very demanding in terms of computing time. The main obstacle is the collapsing formula for the scale hyperparameter of the inverted-Wishart distribution where it is necessary to evaluate a time consuming matrix harmonic mean. Two possibilities for overcoming this difficulty are a square-root implementation (exploiting the Cholesky factorization of the scale hyperparameter) or a Bayesian monitor (West and Harrison, 1986). The latter solution considers the performance of a single routine model in contrast to an alternative model in order to detect a model breakdown. This approach is particularly worth exploring because the detection of a model breakdown may well be the ultimate goal of a multivariate forecasting system. For instance, suppose that the multivariate time series consisted of the vital signs of a hospital patient. A model breakdown would be interpreted, in this case, as a critical change in his/her condition.

Vast forecasting improvements may result from considering a dynamic scale variance matrix; however, although coherence regarding the beliefs about the observations is always assured, it is seldom if ever assured for both observations and parameters. It is possible to construct an evolutional distribution for the scale variance matrix which yields the discount rule given in Chapter 6; unfortunately this is not enough. In order to assure full coherence it is necessary to have an evolutional distribution for both regression and scale variance parameters. Another drawback of discount methods makes itself apparent in long-term forecasting; typically the uncertainty grows in an explosive fashion (Ameen, 1984). Therefore, it may be worthwhile to search for a suitable fully Bayesian formulation of the DLMR with a dynamic scale variance. Meanwhile, the discount method provides the only practical procedure that we know of.

The PIE theory gives sound answers to the problem of employing estimated likelihoods for predictive purposes. In the context of dynamic linear models the use of plug-in estimated likelihoods is recommended for the DWMR model but not for the DLMR as a rule. In particular, for multivariate DLM's the estimated likelihood is incapable of approximating the predictive density. For DWMR models, the predictive multivariate-t distribution can be replaced successfully by a multivariate estimated normal likelihood as the degrees of freedom increase. The PIE for the set of regression parameters is the usual

mean but for the scale variance matrix it is the mean multiplied by a correction factor. This correction factor increases the variance of the estimated distribution compensating for the uncertainty of the regression parameters.

Regarded as point estimators the PIE's have an appealing property: invariance under one-to-one parametric transformations. This makes the PIE's suitable for estimating cumbersome functions of the parameters, e.g. the eigen decomposition of the scale variance matrix in the example of Section 6.4. However, there is a question regarding the consistency of marginal estimators that must be clarified. Consistency is a desirable characteristic of a point estimator, meaning that if the probability density of a parameter is concentrated sequentially around a particular value then a consistent point estimator for the parameter should converge to the same value. Typically, the PIE's associated with a minimal parameterization of a model are consistent as a consequence of the fact that the logarithmic scoring rule is proper. Nevertheless, marginally the PIE's may not be consistent, i.e. the marginal probability density for a subset of parameters may well be concentrating sequentially around a certain point and yet the marginal PIE's may converge to another. For example, the marginal PIE for the scale variance matrix in the DWMR model is not consistent unless the regression parameters are perfectly known. However, consistent marginal PIE's may be enforced simply by demanding that the complementary parameters take the values of their own PIE's, e.g. in the example above the consistent marginal PIE for the scale variance matrix is its mean value. It is worth noting that the use of this consistent marginal PIE has little if any consequence in the discussion of the examples in Sections 5.2 and 6.3. Indeed, the estimated correlations, principal components and their relative importance are unchanged. Notice also that the point estimates in the example of Section 6.2 are implicitly based on these consistent marginal PIE's.

The dynamic recursive model is a very versatile multivariate time series model and yet it is virtually as tractable as its defining submodels. A useful kind of application of the dynamic recursive model is the filtering of information from non-sparse to sparse time series provided that the missing observations conform to a hierarchical scheme. One striking feature of this dynamic model is that it inherits the common facilities of the defining submodels. For instance, if the long-term forecasting distribution for the submodels is obtained easily then it is so for the forecasting distribution of the entire multivariate series. However, this forecasting distribution is given in a conditional form and some marginal forecasting distributions can be, in principle, very difficult to work with, though they may be studied with the help of the simulation technique outlined in Appendix A6.1.

The focus of our attention has been on multivariate dynamic linear normal models. In the case of modelling with DWMR this implies that the feasible region is assumed to be a full high dimensional real space. Apart from this, the most important feasible regions found in practice are the positive orthant and the simplex. These two cases can be handled by means of the multivariate logarithmic and logarithmic ratio transformations which not only restore the feasible region but tend to restore the normality distribution assumptions as well. Moreover, the DWMR (and DLMR) estimators for the

regression parameters and future observations are the best in the linear Bayesian sense regardless of the distribution. However, the problem of non-linearity is very difficult; the usual linearization via Taylor series destroys the DWMR structure in the process. On the other hand, the multivariate extension of the dynamic generalized linear models (West, Harrison and Migon, 1985) is far from trivial; the richness found in the multivariate normal distribution in comparison with other multivariate conjugate priors (e.g. the Dirichlet distribution for multinomial observations) is overwhelming. Therefore, it seems that non-linear non-normal DWMR models are not worth considering, non-linear non-normal dynamic recursive models may very well be.

In summary, the Bayesian forecasting approach for modelling multivariate time series has been very rewarding. Not surprisingly, in our search for answers more questions have been raised and further research is necessary. A seemingly strong criticism of our work could be that virtually no attention has been paid to non-Bayesian methods. We could reply by reproducing here Bayesian arguments regarding the philosophical meaning of probability and the foundations of statistics, but we must confess that the main reason for our choice is, in fact, a mundane one: laziness. Why should we consult a difficult Book of the Horoscope when the Bayesian one is so powerful, clear and easy?

REFERENCES

- (1) Aitchison, J. (1983), "Principal Component Analysis of Compositional Data," *Biometrika*, 70, 1, 57-65.
- (2) Aitchison, J. and Shen, S.M. (1980), "Logistic-normal Distributions: Some Properties and Uses," *Biometrika*, 67, 2, 261-272.
- (3) Amaral, M.A. and Dunsmore, I.R. (1980), "Optimal Estimates of Predictive Distributions," *Biometrika*, 67, 3, 685-689.
- (4) Ameen, J.R.M. (1984), *Discount Bayesian Models and Forecasting*, Ph.D. Thesis, University of Warwick.
- (5) Ameen, J.R.M. and Harrison, P.J. (1984), "Discount Weighted Estimation," *Journal of Forecasting*, 3, 285-296.
- (6) Anderson, B.D.O. and Moore, J.B. (1979), *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, New Jersey.
- (7) Aoki, M. (1967), *Optimization of Stochastic Systems*, Academic Press, New York.
- (8) Bachelier, L. (1900), *Theory of Speculation*, Ph.D. Thesis (translated) in Cootner (1964).
- (9) Barlow, R.E. (1985), "Utility Theory," in *Theory of Reliability*, XCIV Corso Soc. Italiana di Fisica, Bologna.
- (10) Barlow, R.E. (1986), "Influence Diagrams," (prepared for Encyclopedia of Statistics) Bayesian Statistics Study Year 1985/1986 , Department of Statistics, University of Warwick.
- (11) Barnett, V. (1982), *Comparative Statistical Inference* (second edition), Wiley, New York.
- (12) Bayes, T.R. (1763), "An Essay Towards Solving a Problem in the Doctrine of Chances," *Phil. Trans. Roy. Soc. London* 53, 370-418 (reprinted in *Biometrika* (1958), 45, 293-315).
- (13) Berger, J.O. (1985), *Statistical Decision Theory and Bayesian Analysis* (second edition), Springer Verlag, New York.
- (14) Berman, A. and Plemmons, R. (1979), *Non-negative Matrices in the Mathematical Sciences*, Academic Press, New York.
- (15) Bernardo, J.M. (1979), "Expected Information as Expected Utility," *Annals of Statistics*, 7, 3, 686-690.
- (16) Box, G.E.P and Tiao, G.C. (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Massachusetts.
- (17) Broemeling, L.D. (1985), *Bayesian Analysis of Linear Models*, Decker, New York.
- (18) Chen, C.F. (1986), "A Bayesian Approach to Nested Missing-data Problems," in Goel and Zellner (1986).
- (19) Chen, H.F. (1985), *Recursive Estimation and Control for Stochastic Systems*, Wiley, New York.

- (20) Clark, M.R.B. (1981), "The Gauss-Jordan Sweep Operator with Detection of Collinearity," *Applied Statistics*, 31, 166-168.
- (21) Cootner, P.H. (editor), (1964), *The Random Character of the Stock Market Prices*, MIT Press, Cambridge, Massachusetts.
- (22) Dawid, A.P. (1981), "Some Matrix-variate Distribution Theory: Notational Considerations and a Bayesian Application," *Biometrika*, 68, 1, 265-274.
- (23) Dawid, A.P. (1986), "A Bayesian View of Statistical Modelling," in Goel and Zellner (1986).
- (24) de Finetti (1974, 1975), *Theory of Probability* (Vols. 1, 2), Wiley, New York.
- (25) De Groot, M.H. (1970), *Optimal Statistical Decisions*, McGraw-Hill, New York.
- (26) Dempster, A.P. (1969), *Elements of Continuous Multivariate Analysis*, Addison-Wesley, Reading, Massachusetts.
- (27) Dempster, A.P. (1969), "Some Formulas Useful for Covariance Estimation with Gaussian Linear Component Models," in Kallianpur, G., Krishnaiah, P.R. and Ghosh, J.K. (editors), *Statistics and Probability: Essays in Honor of C.R. Rao*, North-Holland, Amsterdam.
- (28) Dempster, A.P. and Carlin, J.B. (1985), Discussion of West, M., Harrison, P.J. and Migon, H.S. (1985).
- (29) Dickey, J.M. (1976), Discussion of Harrison and Stevens (1976).
- (30) Dickey, J.M., Dawid, A.P. and Kadane, J.B. (1986), "Subjective-Probability Assesment Methods for Multivariate-t and Matrix-t Models," in Goel and Zellner (1986).
- (31) Efroymson, M.A. (1960), "Multiple Regression Analysis," in Ralston, A. and Wilf, H.S. (editors), *Mathematical Methods for Digital Computers*, Wiley, New York.
- (32) Fama, E.F. (1965), "The Behaviour of Stock Market Prices," *J. Business*, 38, 34-105.
- (33) Ferguson, T.S. (1972), "A Bayesian Analysis of Some Non-parametric Problems," *Annals of Statistics*, 1, 2, 209-230.
- (34) Gelb, A. (editor), (1974), *Applied Optimal Estimation*, MIT Press (paperback), Cambridge, Massachusetts.
- (35) Goel, P.K. and A. Zellner (editors), (1986), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, North-Holland, Amsterdam.
- (36) Goodnight, J.H. (1979), "A Tutorial on the Sweep Operator," *The American Statistician*, 33, 3, 149-159.
- (37) Granger, C.W.J. (1972), "Emperical Studies of Capital Markets: A Survey," in Szego, G.P. and Shell, K. (editors) *Mathematical Methods in Investment and Finance*, Noth-Holland, Amsterdam.
- (38) Harrison, P.J. and Stevens, C.F. (1971), "A Bayesian Approach for Short-term Forecasting," *Operational Research Quaterly*, 22, 4, 341-362.
- (39) Harrison, P.J. and Stevens, C.F. (1976), "Bayesian Forecasting (with discussion)," *J. R. Statist. Soc.*, B, 38, 205-239.

- Practical* *ed*
- (40) Harrison, P.J. and West, M. (1986), "Bayesian Forecasting in Practice," Invited paper for the 1986 I.O.S. International Conference on Practical Bayesian Statistics, *The Statistician*, (to appear).
 - (41) Hartigan, J.A. (1969), "Linear Bayesian Methods," *J. R. Statist. Soc.*, B, 31, 446-454.
 - (42) Harvey, A.C. (1984), "A Unified View of Statistical Forecasting Procedures," *Journal of Forecasting*, 3, 245-275.
 - (43) Harvey, A.C. (1986), "Analysis and Generalisation of a Multivariate Exponential Smoothing Model," *Management Science*, 32, 2, 374-380.
 - (44) Highfield, R. (1984), "Forecasting with Bayesian State Space Models," Research Report, Graduate School of Business, University of Chicago.
 - (45) Ho, Y.C. and Lee, R.C.K. (1964), "A Bayesian Approach to Problems in Stochastic Estimation and Control," *IEEE Trans. Automat. Control* AC, 9, 5, 333-339.
 - (46) Houle A. (1983), "The Genealogical Tree of Bayesians," Proc. of the 1982 I.O.S. Annual Conference on Practical Bayesian Statistics, *The Statistician*, 32, 1 and 2, 214-215.
 - (47) Jeffreys, H. (1961), *Theory of Probability* (third edition), Clarendon Press, Oxford.
 - (48) Kalman, R.E. (1963), "New Methods in Wiener Filtering," in Bogdanoff, J.L. and Kazen, P. (editors), *Proc. of the First Symposium on Engineering Applications of Random Function Theory and Probability*, Wiley, New York.
 - (49) Kullback, S. and Liebler, R.A. (1951), "On Information and Sufficiency," *Ann. Math. Statist.*, 22, 525-540.
 - (50) Lindley, D.V. (1956), "On the Measure of the Information provided by an Experiment," *Ann. Math. Statist.*, 27, 986-1005.
 - (51) Lindley, D.V. (1971), *Bayesian Statistics, A review*, SIAM, Philadelphia.
 - (52) Marsaglia, G. and Bray, T.A. (1964), "A Convenient Method for Generating Normal Random Variables," *Communications of the Assoc. Comp. Mach.*, 7, 1, 4-10.
 - (53) Maybeck, P.S. (1979, 1982a, 1982b), *Stochastic Models, Estimation and Control* (Vols. 1, 2, 3), Academic Press, New York.
 - (54) Migon, H.S. (1984), *An Approach to Non-linear Bayesian Forecasting Problems with Applications*, Ph.D. Thesis, University of Warwick.
 - (55) Noble, B. and Daniel J.W. (1983), *Applied Linear Algebra*, Prentice-Hall, Englewood Cliffs, New Jersey.
 - (56) O'Hagan, A. (1972), "Some Properties and Applications of the Matrix-normal and Matrix-t Distributions," Manuscript.
 - (57) O'Hagan, A. (1986), "Monte Carlo is Fundamentally Unsound," Proc. of the 1986 I.O.S. International Conference on Practical Bayesian Statistics, *The Statistician*, (to appear).
 - (58) Plackett, R.L. (1950), "Some Theorems in Least Squares," *Biometrika* 37, 149-157.
 - (59) Plackett, R.L. (1972), "Studies in the History of Probability and Statistics XXIX: The Discovery

of the Method of Least Squares," *Biometrika* 59, 2, 239-251.

- (60) Press, S.J. (1982), *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, Krieger, Malabar, Florida.
- (61) Pugachev, V.S. (1965), *Theory of Random Functions and Its Applications to Control Problems* (translated), Pergamon, London.
- (62) Quintana, J.M. (1985a), "Generating Gamma Random Variables with Non-integral Shape Parameters," Research Report, 76, Department of Statistics, University of Warwick.
- (63) Quintana, J.M. (1985b), "A Dynamic Linear Matrix-variate Regression Model," Research Report, 83, Department of Statistics, University of Warwick.
- (64) Quintana, J.M. and West, M. (1986), "An Analysis of International Exchange Rates Using Multivariate D.L.M.'s," Proc. of the 1986 I.O.S. International Conference on Practical Bayesian Statistics, *The Statistician*, (to appear).
- (65) Quintana, J.M., O'Reilly, F.J. and Gomez, S. (1986), "Least Squares with Inequality Restrictions: A Symmetric Positive-definite Linear Complementarity Problem Algorithm," paper submitted to the *Journal of Statistical Computation and Simulation*.
- (66) Raiffa, H. and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Harvard University Press, Cambridge.
- (67) Searle, R.S. (1982), *Matrix Algebra Useful for Statistics*, Wiley, New York.
- (68) Simmons, G.F. (1963), *Introduction to Topology and Modern Analysis*, McGraw-Hill, New York.
- (69) Smith, A.F.M. and West, M. (1983), "Monitoring Renal Transplants; An Application of the Multi-process Kalman Filter," *Biometrics*, 39, 867-878.
- (70) Taylor, S.T. (1980), "Conjectured Models for Trends in Financial Prices, Tests and Forecasts," *J. R. Statist. Soc.*, A, 143, 338-362.
- (71) West, M. (1982), *Aspects of Recursive Bayesian Estimation*, Ph.D. Thesis, University of Nottingham.
- (72) West, M., Harrison, P.J. and Migon, H.S. (1985), "Dynamic Generalized Linear Models and Bayesian Forecasting (with discussion)," *J. A. Statist. Assoc.*, 80, 73-97.
- (73) West, M., and Harrison, P.J. (1986), "Monitoring and adaptation in Bayesian Forecasting Models," *J. A. Statist. Assoc.*, 81, 741-750.
- (74) Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York.