Bayesian Multi- and Matrix-variate Modelling:

Graphical Models and Time Series

by

Hao Wang

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Dr. Mike West, Advisor

_____
Dr. Jerome Reiter

_____
Dr. Sayan Mukherjee

_____
Dr. Tim Bollerslev

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2010

Abstract

(Statistical Science)

Bayesian Multi- and Matrix-variate Modelling: Graphical Models and Time Series

by

Hao Wang

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Dr. Mike West, Advisor

_____
Dr. Jerome Reiter

_____
Dr. Sayan Mukherjee

_____
Dr. Tim Bollerslev

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2010

Abstract

Modelling and inference with higher-dimensional variables, including studies in multivariate time series analysis, raise challenges to our ability to "scale-up" statistical approaches that involve both modelling and computational issues. Modelling issues relate to the interest in parsimony of parametrisation and control over proliferation of parameters; computational issues relate to the basic challenges to the efficiency of statistical computation (simulation and optimisation) with increasingly high-dimensional and structured models. This thesis addresses these questions and explores Bayesian approaches inducing relevant sparsity and structure into parameter spaces, with a particular focus on time series and dynamic modelling.

Chapter 1 introduces the challenge of estimating covariance matrices in multivariate time series problems, and reviews Bayesian treatments of Gaussian graphical models that are useful for estimating covariance matrices. Chapter 2 and 3 introduce the development and application of matrix-variate graphical models and time series models. Chapter 4 develops dynamic graphical models for multivariate financial time series. Chapter 5 and 6 propose an integrated approach for dynamic multivariate regression modelling with simultaneous selection of variables and graphical model structured covariance matrices. Finally, Chapter 7 summarises the dissertation and discusses a number of new and open research directions.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

$x \perp\!\!\!\perp y \mid z$	x and y are conditionally independent given z.
$\text{etr}(\cdot)$	$\exp\{\text{trace}(\cdot)\}$.
$HIW_G(b, \mathbf{\Phi})$	Hyper-inverse Wishart distribution with degree-of-freedom b, location matrix $\mathbf{\Phi}$, and graph G. The density function is given by equation (1.3).
$\mathbf{I}_p$	A $p \times p$ identity matrix.
$N(\mathbf{M}, \mathbf{U}, \mathbf{V})$	Matrix normal distribution with mean $\mathbf{M}$, column and row covariance matrices $\mathbf{U}$ and $\mathbf{V}$ respectively. The density function is given by equation (2.1).
$\mathbf{V} \otimes \mathbf{U}$	Kronecker product of two matrices $\mathbf{U}$ and $\mathbf{V}$.

Abbreviations

CAPM	Capital asset pricing model.
DLM	Dynamic linear model.
FF	Fama-French three-factor model.
i.i.d.	Independent and identically distributed.
MCMC	Markov chain Monte Carlo.
SMC	Sequential Monte Carlo.
SSS	Shotgun stochastic search.
SUR	Seemingly unrelated regression.
SSUR	Sparse seemingly unrelated regression.

Acknowledgements

I am grateful to all people who have helped and inspired me during my doctoral study.

First, I thank my advisor, Mike West. I am very fortunate to have an advisor who is both an international research leader and a terrific mentor. Throughout my graduate study, Mike has been a knowledgeable and inspiring advisor, a bright and engaging colleague, and a caring and trusted friend. Without him, I would have been lost in the challenging transition from a learner to a researcher. He will always be my role model for high standards for excellence in my career.

Special thanks go to Carlos Carvalho, my colleague, friend, and the expert on graphical models and time series. I am also sincerely thankful to Jerry Reiter for opening very interesting and important topics of data confidentiality to me, and patiently guiding me through starting research on that. Further thanks go to David Banks, Tim Bollerslev, David Dunson, Alan Gelfand, Mark Huber, Fan Li, Joe Lucas, Sayan Mukherjee, Scott Schmidler, Dalene Stangl, and Robert Wolpert for their help along the way.

I wish to thank my friends who make my life much sweeter here. Thanks go to Scott Schwartz for restringing my tennis rackets again and again, Guoxian Zhang, Quanlin Li and Kai Cui for "qie cuo" (competition) on Ping Pong, tennis, and poker, Craig Reeson for many enlightening discussions on financial markets, Fei Liu, Liang Zhang, Zhi Ouyang, Huiyan Sang, and Kai Mao for sharing me with tips for

American graduate school surviving, Chunlin Ji for telling me his amazing culinary recipes, and, of course, many others I have left out for helping and supporting me.

I am greatly indebted to my wonderful girlfriend, Sophia Zhengzi Li, who has always been there to encourage me over these years from Tianjin to Durham. Life is better with you standing by me!

Lastly and most importantly, to my parents, Jichang Wang and Jianping Zhu: Thank you for raising me, supporting me, educating me and loving me. To you, I dedicate this work.

1

Introduction

1.1 Bayesian multivariate time series and graphical models

The importance of Bayesian methods in time series analysis has increased rapidly over the last decade. Many application areas are generating time series data of increasing dimension and complexity. Examples include stock returns, changes in labour market employment statistics, or numbers of consumer clicks on displayed web content. Interests often centre on not only forecasting future values but also understanding the relationships between different variables.

One of the most challenging issues in multivariate time series analysis involves the estimation of large-scale covariance matrices. For example, consider the covariance matrix of stock returns. If the returns of Dow Jones 30 companies are of interest, the covariance matrix of these returns at any particular time point t has $30 \times 31/2 = 465$ parameters to be estimated. Even if we can comfortably assume that the N observed returns are i.i.d. with $p \times p$ covariance $\mathbf{\Sigma}$, the eigenstructure of the simple estimator based on the sample covariance matrix tends to be distorted unless p/N is very small. In the case where the temporal dynamics of these covariance matrices are also

of interest, the task of modelling and estimating time-varying covariance matrices becomes more challenging yet very important.

In order to produce stable and robust covariance estimates, modellers must impose structures on the large covariance parameter space. A particularly useful approach to induce structures is based on Gaussian graphical models. This thesis discusses the incorporation of Gaussian graphical model structuring in the general multivariate dynamic environment to improve parameter estimation, structure interpretation and computational efficacy in a variety of important multivariate time series models.

Gaussian graphical modeling offers a powerful set of tools for exploration of multivariate dependence patterns and regularisation of covariance matrices in higher-dimensional problems. I now introduce the basic concepts of the Gaussian graphical models that are the central elements in the approaches developed in the thesis.

1.2 A review of Gaussian graphical models

Gaussian graphical models provide natural tools for modelling conditional independence relationships (Lauritzen, 1996; Whittaker, 1990). They are very useful in many high-dimensional problems: They facilitate computation by decomposing large matrices into many small matrices; they offer stable and robust estimation of covariance by imposing structure; they offer easy understanding of relationship among variables by breaking down a joint probability into a compact representation. This section introduces the essential concepts of graphical models. It starts with the basic representations of graphical models with a focus on undirected graphs. It then reviews the Bayesian framework of Gaussian graphical model analysis; this includes prior specification, posterior and marginal likelihood computation for given graphs, and graphical model uncertainty and search.

1.2.1 Basic graph structure

A graph G is an ordered pair (V, E) where V is a non-empty set of vertices (variables) and E is a set of edges. An edge (a, b) is a *directed* edge from a to b if the ordered pair $(a, b) \in E$ but $(b, a) \notin E$. If an edge $(a, b) \in E$ and also $(b, a) \in E$, then (a, b) is called *undirected* edge. A graph $G = (E, V)$ in which every edge is undirected is called *undirected graph* whereas if every edge is directed the graph is *directed graph*.

Consider an undirected graph $G = (E, V)$ defined by a set of vertices V and a set of edges E. Two vertices a and b are *neighbours* if, and only if the edge $(a, b) \in E$. Consider any subset of nodes $V_A \subseteq V$ and write E_A for the corresponding edge set in G. Then $G_A = (V_A, E_A)$ defines a *subgraph* – an undirected graph on nodes in V_A. Any graph or subgraph is *complete* if all of its vertices are connected by edges in E. A complete graph on p vertices has all $\binom{p}{2}$ edges; otherwise the graph is incomplete. A *clique* is a complete subgraph that is not contained within another complete subgraph. The incomplete graph G can be decomposed into a disjoint triple (G_A, G_B, G_C), if $V_A \cup V_B \cup V_C = V$, and G_C is complete and *separates* G_A and G_B in G (any path from a vertex in V_A to a vertex in V_B goes through V_C). The subgraph G_C is a *separator*. The decomposition is *proper* if $V_A \neq \emptyset$ and $V_B \neq \emptyset$. A sequence of subgraphs that cannot be further properly decomposed are the *prime components* of a graph. A graph is said to be *decomposable* if every prime component is complete so the prime components of a decomposable graph are all cliques. Decomposable graphs have distributional properties that make them particularly tractable, as we shall see in Section 1.2.3.

Any connected graph can be represented as a tree of its prime components – a *junction tree*. In the junction tree, each prime component denoted by P_i is a vertex and if, for any two prime components P_i and P_j and every prime component P_k on the path between them, $P_i \cap P_j \subseteq P_k$. The sets of vertices shared by connected

vertices in the junction tree are called separators of the junction tree, denoted by S_i. In a junction tree, an ordering of the prime components $P_i \in \mathcal{P}$ and separators $S_i \in \mathcal{S}$ as $P_1; S_2, P_2; S_3 P_3; , \cdots$ is said to be *perfect* if, for every $i = 2, 3, \cdots, k$, there exists a $j < i$ such that

$$S_i \quad = \quad P_i \cap H_{i-1} \subset P_j$$

where $H_{i-1} = \bigcup_{j=1}^{i-1} P_j$. Figure 1.1 displays a decomposable graph (top panel) and its junction tree decomposition (bottom panel). For this graph, one perfect ordering for the prime components is: $\{P_1; S_2, P_2; S_3, P_3; S_4, P_4; S_5, P_5\}$, where $P_1 = \{1, 2, 5\}, S_2 = \{2, 5\}, P_2 = \{2, 4, 5, 6\}, S_3 = \{2, 4\}, P_3 = \{2, 3, 4\}, S_4 = \{4, 6\}, P_4 = \{4, 6, 7\}, S_5 = \{6, 7\}$ and $P_5 = \{6, 7, 8, 9\}$.

1.2.2 Graphical models

Probabilistic graphical models rely on graphs that represent independencies among random variables. Each node is a random variable, and a missing edge between two nodes represents conditional independency. Formally, let $\{\mathbf{X}_\nu : \nu \in V\}$ be a collection of random variables indexed by the nodes of an undirected graph $G = (V, E)$. For each $A \subseteq V$, let $\mathbf{X}_A = \{\mathbf{X}_\nu : \nu \in A\}$ indicate the subset of random variables associated with nodes A. A distribution P over V is *Markov* with respect to G if, for any decomposition (A, B) of G, $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_{A \cap B}$. A graphical model is a family of probability distribution which is Markov with respect to a graph.

Decomposition of distributions over graphs

The key to the development of graphical model analysis is the factorisation of a joint distribution that is Markov with respect to a graph G. In general, a p-vector random variable $\mathbf{x} = (x_1, \cdots, x_p)'$ has a multivariate distribution $p(\mathbf{x})$, on a specific graph G that factorises corresponding to the prime components and separators of the junction

tree representation of G, i.e.

$$p(\mathbf{x} \mid G) \quad = \quad \frac{\prod_{P \in \mathcal{P}} p(\mathbf{x}_P)}{\prod_{S \in \mathcal{S}} p(\mathbf{x}_S)} \tag{1.1}$$

where $\mathbf{x}_P$ and $\mathbf{x}_S$ represents the variable subsets on the prime components and separators. The above decomposition is a general and powerful result related to Hammersley and Clifford Theorem (Hammersley & Clifford, 1971). It can be seen that the joint density factors as a product of joint densities of variables within each prime component divided by the product of the joint densities of variables within each separator. By exploring the graph-theoretic representation, the factorisation provides general algorithms for computing marginal and conditional probabilities of interests, and it also controls the computational complexity associated with these problems.

1.2.3 Gaussian graphical modelling

Covariance selection model

A Gaussian graphical model, also known as covariance selection model (Dempster, 1972), is defined by a Gaussian distributed random vector $\mathbf{x} = (x_1, \cdots, x_p)'$ with expectation $\boldsymbol{\mu}$ and non-singular covariance matrix $\boldsymbol{\Sigma}$ (i.e. precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$). In special cases of a multivariate Gaussian distribution on a graph, the conditional dependency property (Markov property) implied by G via its edges is represented by zeros in the precision matrix; that is, for any pair of variables i, j , $\boldsymbol{\Omega}_{ij} = 0$ if and only if $(i, j) \notin E$, which is, by the Markov property, equivalent to $x_i \perp\!\!\!\perp x_j \mid \mathbf{x}_{(1:p)\setminus(i,j)}$. Thus G constrains $\boldsymbol{\Sigma}$ by imposing a pattern of zeros on $\boldsymbol{\Omega}$. In other words, the precision matrix $\boldsymbol{\Omega}$ belongs to $M(G)$, the set of all positive symmetric matrices with elements equal to zeros for any $(i, j) \notin E$. Formal inference identifies graphical structures via patterns of zeros in precision matrix.

Conditional on a graph G and without loss of generality, assume $\boldsymbol{\mu} = \mathbf{0}$. Then

from equation (1.1), the joint distribution of $\mathbf{x}$ is

$$p(\mathbf{x} \mid \mathbf{\Sigma}, G) \;=\; \frac{\prod_{P\in\mathcal{P}} p(\mathbf{x}_P \mid \mathbf{\Sigma}_P)}{\prod_{S\in\mathcal{S}} p(\mathbf{x}_S \mid \mathbf{\Sigma}_S)} \tag{1.2}$$

where, for each $P \in \mathcal{P}$,

$$p(\mathbf{x}_P \mid \mathbf{\Sigma}_P) \;=\; (2\pi)^{-|P|/2}|\mathbf{\Sigma}_P|^{-1/2}\mathrm{etr}\{-\frac{1}{2}\mathbf{x}_P\mathbf{x}_P'\mathbf{\Sigma}_P^{-1}\}$$

and similarly for $p(\mathbf{x}_S \mid \mathbf{\Sigma}_S)$, with $|P|$ denoting the dimension of prime component P. That is, the full joint distribution is completely determined by the component covariance matrices $\mathbf{\Sigma}_P$.

Prior and posteriors for covariance matrices

From a Bayesian perspective, inference on the above covariance selection model involves posterior distribution $p(G, \mathbf{\Sigma} \mid \mathbf{x}) = p(G \mid \mathbf{x})p(\mathbf{\Sigma} \mid \mathbf{x}, G)$ under certain priors $p(G, \mathbf{\Sigma}) = p(\mathbf{\Sigma} \mid G)p(G)$. Hence, two types of uncertainties are associated with the model: uncertainty about the covariance matrix $\mathbf{\Sigma}$ and uncertainty about the graphical structure G. I begin with the prior specification for $\mathbf{\Sigma}$ given a decomposable graph G.

For the parameter $\mathbf{\Sigma}$, the class of *hyper-inverse Wishart* distributions (Dawid & Lauritzen, 1993) for $\mathbf{\Sigma}$ extends the standard multivariate Gaussian/inverse-Wishart framework to graphs. The prior for $\mathbf{\Sigma}$, $p(\mathbf{\Sigma} \mid G)$, is hyper-inverse Wishart, $HIW_G(\delta, \mathbf{\Phi})$ with degree-of-freedom parameter δ and location matrix $\mathbf{\Phi} > 0$. The prior density factors in a form related to the likelihood (1.2), namely

$$p(\mathbf{\Sigma} \mid G) \;=\; \frac{\prod_{P\in\mathcal{P}} p(\mathbf{\Sigma}_P \mid \delta, \mathbf{\Phi}_P)}{\prod_{S\in\mathcal{S}} p(\mathbf{\Sigma}_S \mid \delta, \mathbf{\Phi}_S)} \tag{1.3}$$

where, for each complete prime component $P \in \mathcal{P}$, the corresponding sub-matrix

$\mathbf{\Sigma}_P$ has an inverse Wishart $IW(\delta, \mathbf{\Phi}_P)$ prior with density

$$p(\mathbf{\Sigma}_P \mid G) = \frac{|\frac{\mathbf{\Phi}_P}{2}|^{(\frac{\delta+|P|-1}{2})}}{\Gamma_{|P|}(\frac{\delta+|P|-1}{2})} |\mathbf{\Sigma}_P|^{-\frac{\delta+2|P|}{2}} \text{etr}\{-\frac{1}{2}\mathbf{\Phi}_P \mathbf{\Sigma}_P^{-1}\} \tag{1.4}$$

where $\Gamma_k(a)$ is the multivariate gamma function $\Gamma_k(a) = \pi^{\frac{k(k-1)}{4}} \prod_{i=0}^{k-1} \Gamma(a - i/2)$. Since decomposable graphs consist entirely of complete prime components, equations (1.3) and (1.4) indicate that the submatrices of $\mathbf{\Phi}$ corresponding to the variables $P \in \mathcal{P}$ determine the prior. Any graph G determines which collection of submatrices of $\mathbf{\Phi}$ are to be taken to form a hyper-inverse Wishart prior on $\mathbf{\Sigma}$.

This family of hyper-inverse Wishart priors $HIW_G(\delta, \mathbf{\Phi})$ over different decomposable graphs is a general class of priors that have the desired properties as a prior for quantities parameterising a graphical model:

- It is compatible in the sense that, for two graphs G_1 and G_2 and for any clique A common to both G_1 and G_2, the marginal priors on A induced by $HIW_{G_1}(\delta, \mathbf{\Phi})$ and $HIW_{G_2}(\delta, \mathbf{\Phi})$ are the same.

- It is consistent in the sense that, for any decomposition (A, B) of G, $p(\mathbf{\Sigma}_A \mid G)$ and $p(\mathbf{\Sigma}_B \mid G)$ induce the same prior distribution over $\mathbf{\Sigma}_{A \cap B}$.

The hyper-inverse Wishart prior is conjugate for any decomposable graph G (Dawid & Lauritzen, 1993). That is, for a multivariate Gaussian random sample $\mathbf{X} = (\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)})$ of size n, if $p(\mathbf{\Sigma} \mid G) = HIW_G(\delta, \mathbf{\Phi})$, the posterior is hyper-inverse Wishart $\mathbf{\Sigma} \mid \mathbf{X} \sim HIW_G(\delta + n, \mathbf{\Phi} + \mathbf{S_X})$ with $\mathbf{S_X}$ the sum of products matrix $\sum_{i=1}^{n} \mathbf{x}^{(i)} \mathbf{x}^{(i)'}$. The marginal distribution for $\mathbf{X}$ has a *hyper-t* distribution denoted by $HT_G(0, \delta, \mathbf{\Phi})$, extending the standard matrix t distribution.

Marginal likelihood for given graphs

The marginal likelihood function for any graph G is computed by integrating out the covariance matrix with respect to the prior,

$$p(\mathbf{X} \mid G) \quad = \quad \int_{\mathbf{\Sigma}^{-1}=\mathbf{\Omega}\in M(G)} p(\mathbf{X} \mid \mathbf{\Sigma}, G)p(\mathbf{\Sigma} \mid G)\mathrm{d}\mathbf{\Sigma}$$

where $M(G)$, as before, indicates the set of all positive-definite symmetric matrices spaces constrained by G.

Under a hyper-inverse Wishart prior for $\mathbf{\Sigma}$ and observed data $\mathbf{X}$ of sample size n, the above integration for the decomposable graph becomes a simple function of the prior and posterior normalising constants, $H(\delta, \mathbf{\Phi}, G)$ and $H(\delta + n, \mathbf{\Phi} + \mathbf{S}_{\mathbf{X}}, G)$:

$$p(\mathbf{X} \mid G) \quad = \quad (2\pi)^{-np/2}\frac{H(\delta, \mathbf{\Phi}, G)}{H(\delta + n, \mathbf{\Phi} + \mathbf{S}_{\mathbf{X}}, G)}$$

where the normalising constant $H(\delta, \mathbf{\Phi}, G)$ is given by

$$H(\delta, \mathbf{\Phi}, G) \quad = \quad \frac{\prod_{P\in\mathcal{P}} |\frac{\mathbf{\Phi}_P}{2}|^{(\frac{\delta+|P|-1}{2})}\Gamma_{|P|}(\frac{\delta+|P|-1}{2})^{-1}}{\prod_{S\in\mathcal{S}} |\frac{\mathbf{\Phi}_S}{2}|^{(\frac{\delta+|S|-1}{2})}\Gamma_{|S|}(\frac{\delta+|S|-1}{2})^{-1}}.$$

Priors over graphs

In Section 5.2.3, we model the uncertainty about the covariance matrix $\mathbf{\Sigma}$ by introducing a conjugate prior over $\mathbf{\Sigma}$ for any fixed graph structure. I now define the priors over graph space to model the uncertainty over graph structures G.

I first define edge inclusion indicators e_{ij} of a graph G as follows:

$$e_{ij} = \begin{cases} 1, & \text{if } (i,j) \in E \\ 0, & \text{otherwise.} \end{cases}$$

An independent Bernoulli prior with parameter β on each edge inclusion probability is used as an initial sparse inducing prior. That is, a p node graph $G = (E, V)$ with

$|E|$ edges and $T = p(p-1)/2$ possible edges has prior probability

$$\begin{aligned} p(G) &\propto \prod_{i,j} \beta^{e_{ij}} (1-\beta)^{1-e_{ij}} \\ &= \beta^{|E|} (1-\beta)^{T-|E|}. \end{aligned}$$

This prior distribution has its mode at $T \times \beta$ edges for an unrestricted (decomposable and nondecomposable) p node graph. Therefore, if $\beta = 1/2$, i.e. a uniform prior over all graphs, the prior favors models in which the number of edges is quite large, that is, a set of modal graphs with $T/2 = p(p-1)/4$ nodes per graph. To induce sparsity and hence obtain parsimonious representation of the conditional independence structure implied by a graph, we prefer a much smaller β, for example $\beta = 2/(p-1)$, which, in unrestricted cases, would provide a prior mode at p edges. For decomposable graph space, Carvalho (2006) explores how an appropriate β could offer sparse graph priors in comparison with uniform priors.

Central to this analysis is the edge inclusion probability parameter β – a critical sparsity inducing parameter. Learning about β from the data can be addressed with a view to embedding β into the MCMC sampling analysis. For example, using the conjugate beta priors for β, we can draw the posterior samples of β given other parameters from a beta distributions. Another approach to dealing with the uncertainty about β is that we can also fully integrate out β with respect to its beta prior; we discuss such approach in Section 5.3.3.

The above approach to prior specification penalises the number of edges. One could also penalise other measure of model complexity such as the maximum or average prime component size, number of cliques, etc (see for example Wong et al. 2003).

1.2.4 Local updates for decomposable graphs

In addition to the analytical expressions for marginal likelihoods, decomposable graphs have attractive properties that are related to local updates. In particular, for any two decomposable graphs G and G' that differ by one edge only, calculating their marginal likelihood ratio $p(\mathbf{X} \mid G)/p(\mathbf{X} \mid G')$ is facilitated by the similarity between G and G' in a sense that the computation requires far less than the complete calculation of two likelihoods. Giudici & Green (1999) first exploited this property. Wong et al. (2003) further simplified that computation by examining the Cholesky decomposition of the component subgraph that contains the differing edge. In contrast, non-decomposable graphs do not share this attractive property because a single edge change may thoroughly change the junction tree representation of components.

In the dynamic financial index models in Chapter 4, we exploit this property in our sequential stochastic search methods for exploring decomposable graphical model uncertainty. In the other two models proposed in this thesis, although we focus on the decomposable graphs, we still cannot take advantage of the property. The marginal likelihoods for these two models do not have the analytical form and cannot factor over prime components.

1.2.5 Graphical model search algorithms

The previous sections have described how to analytically evaluate the unnormalised posterior probability $p(G) \propto p(\mathbf{X} \mid G)p(G)$ for any decomposable graph G. The remaining substantial problem is how to effectively explore this posterior graphical model space. These model spaces are usually enormous; p nodes in a graph mean $T = p(p-1)/2$ possible edges, and hence 2^T possible graphs corresponding to all combinations of individual edges being in or out of the graph. Even for decomposable graphs and moderate p, it is impossible to enumerate all possible graphs in the space. This motivates the development of effective search algorithm for exploring graphical

model uncertainty. In order to be accurate and scalable, it is desirable for these search algorithms to be able to quickly move towards high posterior probability region, and also to take advantage of local computation. Two classes of stochastic search algorithms have been developed in the literature as we now describe below.

Markov chain Monte Carlo algorithms

MCMC is widely used to assist Bayesian treatments of model selection, and account for model uncertainty for discrete graphical models (Madigan & York, 1995; Dellaportas & Forster, 1999; Giudici & Castelo, 2003). For Gaussian graphical models, Wong et al. (2003) use their results to construct a fixed scan Gibbs sampler for decomposable graphs and Giudici & Green (1999) implement a reversible jump Markov chain Monte Carlo sampler to carry out the graphical model determination. With decomposable graphical models and conjugate priors, explicit marginal likelihood are available, allowing the use Metropolis-Hastings sampler to stochastically explore the model space. In general, starting from a current graph $G = (V, E)$, a candidate graph $G' = (V, E')$ is proposed from a proposal distribution $q(G'; G)$ and accepted with probability

$$\alpha \quad = \quad min\left\{\frac{p(G' \mid \mathbf{X})q(G; G')}{p(G \mid \mathbf{X})q(G'; G)}, 1\right\}.$$

There are several ways to construct the proposal: fixed scan Gibbs, Metropolis-Hastings in which the edge to be updated is picked up at random, and Metropolis-Hastings in which the choice to add or delete an edge is made, and then an edge was selected at random from those appropriate for that type of move. As is noted by Jones et al. (2005), there is no noticeable difference in performance between these closely related MCMC algorithms.

Shotgun stochastic search

Shotgun stochastic search (SSS) (Jones et al., 2005; Hans, 2005) is a stochastic search algorithm that is inspired by MCMC methods but differs with MCMC in its key innovation: SSS takes advantage of distributed computing environment to parallelism computations and to more rapidly identify promising models. The SSS algorithm can be summarised as:

i. Start with a graph G.

ii. Define a neighbourhood of proposal graphs by randomly choosing F_1 neighbours of G.

iii. Evaluate the posterior probability of graphs (up to a normalising constant) in the neighbourhood in parallel, and retain the top F_2 graphs.

iv. Choose a new current model from the top F_2 graphs with each G_i having probability proportional to $p(G_i \mid \mathbf{X})^\alpha$. Here $\alpha > 0$ is an annealing parameter.

v. Go to step (ii) and iterate.

In Step (ii), the neighbourhood of the current graph must be sufficiently comprehensive to allow the search to move easily throughout the graphical model space. This is usually accomplished by considering each possible edge change in one of the proposal models at each iteration. Step (iii) has the most computational burden – a problem that can be solved by parallelised computation.

1.3 Thesis outline

Chapter 2 and 3 introduce the development and application of matrix-variate graphical models and time series models. This develops a complete Bayesian analysis of matrix normal graphical models, i.e., matrix normal distributions in which each

of the characterising variance matrices is constrained by a set of conditional independence restrictions consistent with an underlying graphical model. Part of the motivation lies in the interest in scaling matrix-variate models to deal with increasingly higher-dimensional problems, such as multiple economic indicators or assets measured across multiple funds, companies, sectors or countries.

Chapter 4 develops dynamic graphical models for multivariate financial time series. The method described in this chapter is aimed to flexibly yet tractably forecast large-scale covariance matrices. One theoretical advantage of our methods is that the assumption of uncorrelated residuals in popular financial index models has been relaxed. Moreover, it will be shown empirically that the synthesis of dynamic graphical models with financial index models generally improves the covariance matrix forecasting relative to standard financial index models.

Chapters 5 and 6 propose an integrated approach for seemingly unrelated regression (SUR) modelling with simultaneous selection of variables and graphical-model structured covariance matrices. Current developments of SUR are almost all based on non-sparse modelling ideas. These methods are useful when the number of regressions and predictors is small. To scale up the SUR analysis to higher-dimensional problems, Chapter 5 combines variable selections, graphical models and SUR to produce an effective method for sparse SUR (SSUR) inferences. The new SSUR is then compared with SUR in carefully designed experiments in Chapter 5. In these experiments, SSUR is shown to both generate relevant structures, and outperform SUR in terms of risks of parameter estimations. Based on the development of SSUR in Chapter 5, Chapter 6 then describes a detailed application of SSUR to mutual fund performance evaluations, followed by an extension SSUR to more general dynamic models.

Finally, Chapter 7 summarises the dissertation and discusses a number of new and open research directions.

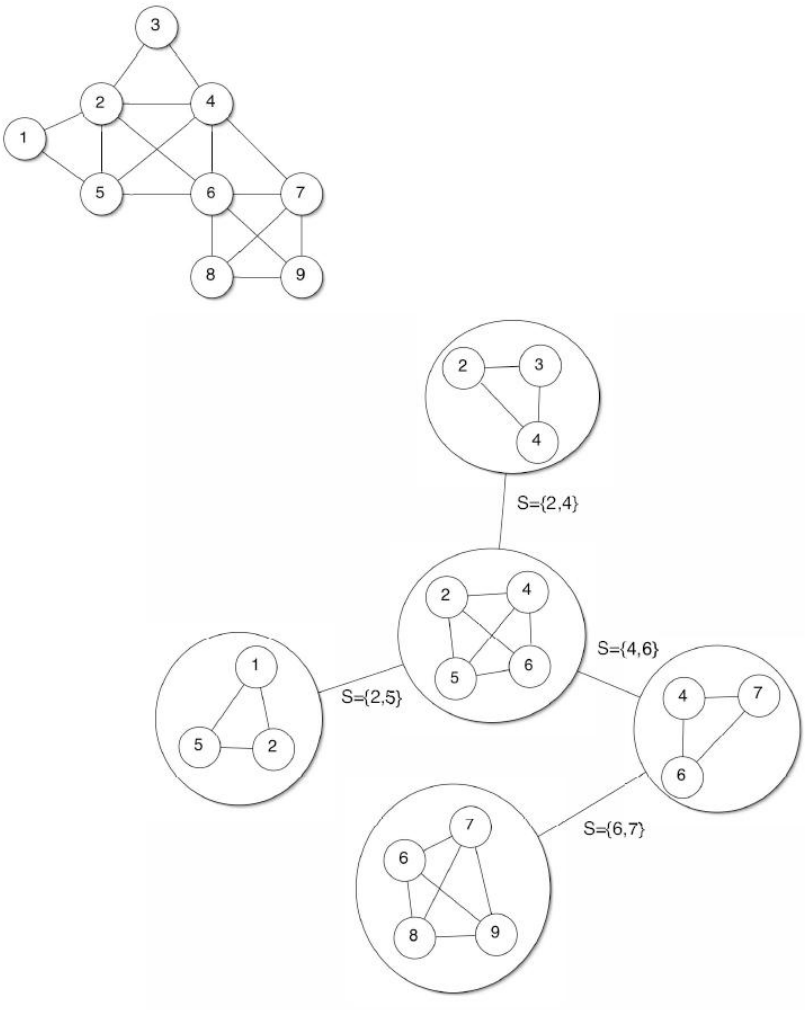

FIGURE 1.1: A decomposable graph and its junction tree decomposition (Carvalho, 2006). Here one perfect ordering of the prime components is $\{P_1; S_2, P_2; S_3, P_3; S_4, P_4; S_5, P_5\}$, where $P_1 = \{1, 2, 5\}, S_2 = \{2, 5\}, P_2 = \{2, 4, 5, 6\}, S_3 = \{2, 4\}, P_3 = \{2, 3, 4\}, S_4 = \{4, 6\}, P_4 = \{4, 6, 7\}, S_5 = \{6, 7\}$ and $P_5 = \{6, 7, 8, 9\}$.

2

Matrix normal graphical models

2.1 Introduction

Matrix-variate normal distributions (Dawid, 1981; Gupta & Nagar, 2000) have been studied in analysis of two-factor linear models for cross-classified multivariate data (Finn, 1974; Galecki, 1994; Naik & Rao, 2001), in spatio-temporal models (Mardia & Goodall, 1993; Huizenga et al., 2002) and other areas. Some computational and inferential techniques, including iterative calculation of maximum likelihood estimates have been developed (Dutilleul, 1999; Mitchell et al., 2005, 2006), and some empirical Bayesian methodology has recently been introduced for Procrustes analysis with matrix models (Theobald & Wuttke, 2006).

This chapter together with Chapter 3 develops a complete Bayesian analysis of matrix normal graphical models, i.e., matrix normal distributions in which each of the characterising variance matrices is constrained by a set of conditional independence restrictions consistent with an underlying graphical model (Whittaker, 1990; Lauritzen, 1996; Giudici, 1996; Giudici & Green, 1999; Jones et al., 2005). The framework includes fully Bayesian analysis of the matrix normal (full graphs) as a special

case, and effective computational methods for marginal likelihood computation on a specified graphical model that underlies inference about conditional independence structures. The developments include novel Markov random field models, with potential utility in spatial and image analysis, that emerge naturally as a sub-class of matrix normal graphical models. The framework then extends the random sampling framework to matrix-variate time series models that inherit the graphical model structure to represent conditional independencies in matrix series over time.

This chapter focuses on matrix normal graphical models for given graphs. It begins with preliminaries and notation for matrix normal models in Section 2.2, followed by elements of Bayesian analysis of this class of models for fixed graphs. Section 2.5 provides a simple example of analysis of a simulated data set, illustrating aspects of the computation. Section 2.6 shows how the matrix graphical structure can be naturally embedded in a broad class of matrix time series models, and develops a detailed analysis of a macro-economic data set for additional illustration of the effectiveness and utility of the new matrix-variate models.

2.2 Matrix variate normals, graphs and notation

The $q \times p$ random matrix $\mathbf{Y}$ is matrix normal, $\mathbf{Y} \sim N(\mathbf{M}, \mathbf{U}, \mathbf{V})$, with mean $\mathbf{M}$ $(q \times p)$, column and row covariance matrices $\mathbf{U} = (u_{ij})$ $(q \times q)$ and $\mathbf{V} = (v_{ij})$ $(p \times p)$ respectively, when

$$p(\mathbf{Y}) \equiv p(\mathbf{Y} \mid \mathbf{U}, \mathbf{V}) = k(\mathbf{U}, \mathbf{V}) \exp[-\mathrm{tr}\{(\mathbf{Y} - \mathbf{M})'\mathbf{U}^{-1}(\mathbf{Y} - \mathbf{M})\mathbf{V}^{-1}/2\}] \qquad (2.1)$$

where $k(\mathbf{U}, \mathbf{V}) = (2\pi)^{-qp/2}|\mathbf{U}|^{-p/2}|\mathbf{V}|^{-q/2}$. The rows $\mathbf{y}_{i\star}$, $(i = 1, \ldots, p)$, and columns $\mathbf{y}_{\star j}$, $(j = 1, \ldots, q)$, have margins $\mathbf{y}_{i\star} \sim N(\mathbf{m}_{i\star}, u_{ii}\mathbf{V})$ and $\mathbf{y}_{\star j} \sim N(\mathbf{m}_{\star j}, v_{jj}\mathbf{U})$ with precisions $\mathbf{\Omega} = \mathbf{U}^{-1} = (\omega_{ij})$ and $\mathbf{\Lambda} = \mathbf{V}^{-1} = (\lambda_{ij})$. The normal conditional distribu-

tions have mean vectors and covariance matrices given by

$$E(\mathbf{y}_{i\star} \mid \mathbf{y}_{-i\star}) = \mathbf{m}_{i\star} - \omega_{ii}^{-1} \textstyle\sum_{s\in(1,\ldots,q\setminus i)} \omega_{is}(\mathbf{y}_{s\star} - \mathbf{m}_{s\star}), \quad \mathrm{cov}(\mathbf{y}_{\mathrm{i}\star} \mid \mathbf{y}_{-\mathrm{i}\star}) = \omega_{ii}^{-1}\mathbf{V},$$
$$E(\mathbf{y}_{\star j} \mid \mathbf{y}_{-\star j}) = \mathbf{m}_{\star j} - \lambda_{jj}^{-1} \textstyle\sum_{t\in(1,\ldots,p\setminus j)} \lambda_{tj}(\mathbf{y}_{\star t} - \mathbf{m}_{\star t}), \quad \mathrm{cov}(\mathbf{y}_{\star\mathrm{j}} \mid \mathbf{y}_{-\star\mathrm{j}}) = \lambda_{jj}^{-1}\mathbf{U},$$

for rows $i = 1, \ldots, q$ and columns $j = 1, \ldots, p$. Zeros in $\mathbf{\Omega}$ and/or $\mathbf{\Lambda}$ define conditional independencies. If $(i, j) \neq (s, t)$ then y_{ij} and y_{st} may, conditional upon $y_{-(ij,st)}$ be dependent through either rows or columns; conditional independence is equivalent to: (a) at least one zero among λ_{tj} and ω_{is} when $s \neq i, j \neq t$; (b) $\omega_{is} = 0$ when $s \neq i, j = t$; (c) $\lambda_{jt} = 0$ when $s = i, j \neq t$. With no loss of generality in this section I set $\mathbf{M} = \mathbf{0}$.

Undirected graphical models can be applied to each of $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ to represent strict conditional independencies. A graph $G_{\mathbf{V}}$ on nodes $\{1, \ldots, p\}$ has edges between pairs of column indices (j, t) for which $\lambda_{jt} \neq 0$; $\mathbf{\Lambda}$ has off-diagonal zeros corresponding to within-row conditional independencies. Similarly, a graph $G_{\mathbf{U}}$ on nodes $\{1, \ldots, q\}$ lacks edges between row indices (i, s) for which $\omega_{is} = 0$. I focus here on decomposable graphs $G_{\mathbf{U}}$ and $G_{\mathbf{V}}$. The theory of graphical models can be now overlaid to define conditional factorisations of the matrix normal density over graphs. Over $G_{\mathbf{V}}$, for example, I have

$$p(\mathbf{Y} \mid \mathbf{U}, \mathbf{V}, G_{\mathbf{V}}, G_{\mathbf{U}}) = \frac{\prod_{P_{\mathbf{V}}\in\mathcal{P}_{\mathbf{V}}} p(\mathbf{Y}_{\star P_{\mathbf{V}}} \mid \mathbf{U}, \mathbf{V}_{P_{\mathbf{V}}})}{\prod_{S_{\mathbf{V}}\in\mathcal{S}_{\mathbf{V}}} p(\mathbf{Y}_{\star S_{\mathbf{V}}} \mid \mathbf{U}, \mathbf{V}_{S_{\mathbf{V}}})} \tag{2.2}$$

where $\mathcal{P}_{\mathbf{V}}$ is the set of complete prime components, or cliques, of $G_{\mathbf{V}}$ and $\mathcal{S}_{\mathbf{V}}$ is the set of separators. For each subgraph $g \in \{\mathcal{P}_{\mathbf{V}}, \mathcal{S}_{\mathbf{V}}\}$, $\mathbf{Y}_{\star g}$ is the $q \times |g|$ matrix with variables from the $|g|$ columns of $\mathbf{Y}$ defined by the subgraph, and $\mathbf{V}_g$ the corresponding sub-matrix of $\mathbf{V}$. Each term in equation (5.3) is matrix normal, $\mathbf{Y}_{\star g} \sim N(\mathbf{0}, \mathbf{U}, \mathbf{V}_g)$ with $\mathbf{\Lambda}_g = \mathbf{V}_g^{-1}$ having no off-diagonal zeros. I can similarly factorise the joint density over $G_{\mathbf{U}}$.

Now, $\mathbf{U}$ and $\mathbf{V}$ are not uniquely identified since, for any $c \neq 0$, $p(\mathbf{Y} \mid \mathbf{U}, \mathbf{V}) = p(\mathbf{Y} \mid c\mathbf{U}, \mathbf{V}/c)$. There are a number of potential approaches to imposing identification constraints such as $\text{tr}(\mathbf{V}) = p$ (Theobald & Wuttke, 2006), and possible strategies that use unconstrained parameters; I discuss the latter in Section 3.5. Our use of hyper-Markov priors over each of $\mathbf{U}$ and $\mathbf{V}$ with underlying graphical models, discussed below, makes it desirable to adopt an explicit constraint and I enforce $v_{11} = 1$ from here on.

2.3 Matrix graphical modelling

In Section 1.2.3, I have introduced the hyper-inverse Wishart priors of equation (1.3) that are conjugate for covariance matrices in multivariate normal graphical models (Dawid & Lauritzen, 1993). Hyper-inverse Wishart distributions are compatible and consistent across graphs, which is critical when admitting uncertainty about graph structures (Giudici & Green, 1999; Jones et al., 2005). On decomposable graphs, the implied priors on sub-covariance matrices of all components and separators are inverse Wishart. Use of independent hyper-inverse Wishart priors for $\mathbf{U}, \mathbf{V}$ in the current context is a natural choice, and maintains compatibility and consistency across graphs $G_{\mathbf{U}}, G_{\mathbf{V}}$. To incorporate the identification constraint $v_{11} = 1$, I use a parameter expansion approach. Parameter expansion involves expanding the parameter space by adding new nuisance parameters, and has been used simply algorithmically to accelerate Markov chain Monte Carlo samplers (Liu et al., 1998; Liu & Wu, 1999), but can also be used to induce new priors (Gelman, 2004, 2006) as is germane here.

I assume the prior $p(\mathbf{U}, \mathbf{V}) = p(\mathbf{U})p(\mathbf{V})$ where the margins are defined by

$$\mathbf{U} \sim HIW_{G_{\mathbf{U}}}(b, \mathbf{B}) \quad \text{and} \quad \mathbf{V} = \mathbf{V}^*/v_{11}^* \quad \text{where} \quad \mathbf{V}^* \sim HIW_{G_{\mathbf{V}}}(d, \mathbf{D}). \tag{2.3}$$

The parameter expansion concept relates to v_{11}^* as an added parameter that converts column scales in $\mathbf{V}$ to those relative to the scale of the first column. As

I move across graphs $G_{\mathbf{V}}$, the priors $p(\mathbf{V} \mid G_{\mathbf{V}})$ have the same induced priors over subgraph correlation structures although are no longer in complete agreement for $\mathbf{V} = \mathbf{V}^*/v_{11}^*$ due to the different parameterisations and interpretations. This is natural and appropriate. Suppose $G_{\mathbf{V}}$ and $G'_{\mathbf{V}}$ are two graphs with a common clique C. Each element in $\text{diag}(\mathbf{V}_C)$ represents the relative scale of variance of that column to the variance of the first column so that, if $G_{\mathbf{V}}$ and $G'_{\mathbf{V}}$ imply different conditional dependencies between the first column and columns linked to C, then the induced priors over $\mathbf{V}_C$ should indeed be different.

The prior $p(\mathbf{V})$ is obtained by transformation from $\mathbf{V}^*$. On any graph $G_{\mathbf{V}}$, $\mathbf{V}$ is determined only by those free elements appearing in the sub-matrices corresponding to the cliques of the graph, and the non-free elements of $\mathbf{V}$ are deterministic functions of the free elements (Carvalho et al., 2007). Let ν be the number of free elements; then the transformation from $\mathbf{V}^*$ to $(\mathbf{V}, v_{11}^*)$ has Jacobian $(v_{11}^*)^{\nu-1}$ leading to

$$p(\mathbf{V}, v_{11}^*) \;=\; HIW_{G_{\mathbf{V}}}(v_{11}^*\mathbf{V} \mid d, \mathbf{D})(v_{11}^*)^{\nu-1}. \tag{2.4}$$

Coupled with the prior $p(\mathbf{U})$ on $G_{\mathbf{U}}$, this defines a class of conditionally conjugate priors in the expanded parameter space.

2.4 Gibbs sampling on given graphs

Assume an initial random sampling context with $q \times p$ data matrices $\mathbf{Y}_i, (i = 1, \ldots, n)$, drawn independently from equation (2.1), and write $\mathbf{Y}$ for the full set of data. It is easy to see that, on specified graphs $(G_{\mathbf{U}}, G_{\mathbf{V}})$, the posterior simulation of $p(\mathbf{U}, \mathbf{V}, v_{11}^* \mid \mathbf{Y})$ can be implemented in a Gibbs sampler format, a complete sweep of which consists of the following three steps:

i. Sampling $p(\mathbf{U} \mid -)$:

The posterior distribution of $\mathbf{U}$ takes the form

$$p(\mathbf{U} \mid \mathbf{Y}_{1:n}, \mathbf{V}, v_{11}^*) \;\propto\; p(\mathbf{Y}_{1:n} \mid \mathbf{V}, v_{11}^*, \mathbf{U})p(\mathbf{U} \mid G_{\mathbf{U}})$$

which shows that the update of $\mathbf{U}$ can be carried out with the standard conjugacy of normal and hyper-inverse Wishart for any graph $G_{\mathbf{U}}$ giving the following posterior conditional distribution of $\mathbf{U}$:

$$(\mathbf{U} \mid \mathbf{Y}_{1:n}, \mathbf{V}, v_{11}^*) \sim HIW_{G_{\mathbf{U}}}(b + np, \mathbf{B} + \sum_{i=1}^{n} \mathbf{Y}_i \mathbf{V}^{-1} \mathbf{Y}_i). \tag{2.5}$$

ii. Sampling $p(v_{11}^* \mid -)$:

Notice that the joint prior distribution of $(\mathbf{V}, v_{11}^*)$ has the form as in equation (2.4). So, the conditional posterior for v_{11}^* is given by

$$\begin{aligned} p(v_{11}^* \mid \mathbf{Y}_{1:n}, \mathbf{V}, \mathbf{U}) &\propto p(v_{11}^* \mid \mathbf{V}) \\ &\propto \frac{\prod_{P_{\mathbf{V}} \in \mathcal{P}_{\mathbf{V}}} p((v_{11}^* \mathbf{V})_{P_{\mathbf{V}}} \mid d, \mathbf{D}_{P_{\mathbf{V}}})}{\prod_{S_{\mathbf{V}} \in \mathcal{S}_{\mathbf{V}}} p((v_{11}^* \mathbf{V})_{S_{\mathbf{V}}} \mid d, \mathbf{D}_{S_{\mathbf{V}}})} (v_{11}^*)^{\nu - 1}, \end{aligned}$$

where the distribution for each prime component (and separator) is an inverse Wishart. Hence the density for each $P_{\mathbf{V}} \in \mathcal{P}_{\mathbf{V}}$ can be written as

$$\begin{aligned} p((v_{11}^* \mathbf{V})_{P_{\mathbf{V}}} \mid d, \mathbf{D}_{P_{\mathbf{V}}}) &\propto |(v_{11}^* \mathbf{V})_{P_{\mathbf{V}}}|^{-(d+2|P_{\mathbf{V}}|)/2} \mathrm{etr}\{-\frac{1}{2}(v_{11}^* \mathbf{V})_{P_{\mathbf{V}}}^{-1} \mathbf{D}_{P_{\mathbf{V}}}\} \\ &\propto (v_{11}^*)^{-|P_{\mathbf{V}}|(2|P_{\mathbf{V}}|+d)/2} \mathrm{etr}\{-\frac{1}{2}(v_{11}^*)^{-1} (\mathbf{V}_{P_{\mathbf{V}}}^{-1} \mathbf{D}_{P_{\mathbf{V}}})\} \end{aligned}$$

yielding

$$\begin{aligned} p(v_{11}^* \mid \mathbf{Y}_{1:n}, \mathbf{V}, \mathbf{U}) \propto (v_{11}^*)^{-(a/2-\nu)-1} \mathrm{etr}\{-\frac{1}{2}(v_{11}^*)^{-1} \big(\sum_{P_{\mathbf{V}} \in \mathcal{P}_{\mathbf{V}}} \mathbf{V}_{P_{\mathbf{V}}}^{-1} \mathbf{D}_{P_{\mathbf{V}}} - \\ - \sum_{S_{\mathbf{V}} \in \mathcal{S}_{\mathbf{V}}} \mathbf{V}_{S_{\mathbf{V}}}^{-1} \mathbf{D}_{S_{\mathbf{V}}} \big)\} \end{aligned}$$

where $a = \sum_{P_{\mathbf{V}}} |P_{\mathbf{V}}|(2|P_{\mathbf{V}}| + d) - \sum_{S_{\mathbf{V}}} |S_{\mathbf{V}}|(2|S_{\mathbf{V}}| + d)$.

Now, note that $\mathbf{V}^{-1}$ can be expressed as (Lauritzen, 1996)

$$\mathbf{V}^{-1} = \sum_{P_{\mathbf{V}} \in \mathcal{P}_{\mathbf{V}}} [\mathbf{V}_{P_{\mathbf{V}}}^{-1}]^0 - \sum_{S_{\mathbf{V}} \in \mathcal{S}_{\mathbf{V}}} [\mathbf{V}_{S_{\mathbf{V}}}^{-1}]^0$$

where K^0 denotes an extension of the matrix K with zeros so as to give it the appropriate dimensions; then the density function of v_{11}^* can be further simplified as

$$p(v_{11}^* \mid \mathbf{Y}_{1:n}, \mathbf{V}, \mathbf{U}) \quad \propto \quad (v_{11}^*)^{-(a/2-\nu)-1} \text{etr}\{-\frac{1}{2}(v_{11}^*)^{-1}\mathbf{D}\mathbf{V}^{-1}\}$$

which implies

$$(v_{11}^* \mid \mathbf{Y}_{1:n}, \mathbf{V}, \mathbf{U}) \quad \sim \quad IG(a/2-\nu, \text{tr}(\mathbf{D}\mathbf{V}^{-1})/2). \tag{2.6}$$

iii. Sampling $p(\mathbf{V} \mid -)$:

The complete conditional posterior distribution for $\mathbf{V}$ satisfying $v_{11} = 1$ can be expressed as

$$p(\mathbf{V} \mid \mathbf{Y}_{1:n}, \mathbf{U}, v_{11}^*) \quad \propto \quad p(\mathbf{Y}_{1:n} \mid \mathbf{V}, v_{11}^*, \mathbf{U}) p(\mathbf{V} \mid G_{\mathbf{V}}, v_{11}^*)$$

where the prior $p(\mathbf{V} \mid G_{\mathbf{V}}, v_{11}^*)$, from equation (2.4), is

$$HIW_{G_{\mathbf{V}}}(\mathbf{V} \mid d, \mathbf{D}v_{11}^{*-1}) I(\mathbf{V}_{11} = 1).$$

Therefore, by the same argument as in i., I obtain the conditional posterior distribution for $\mathbf{V}$ as

$$(\mathbf{V} \mid -) \sim \quad HIW_{G_{\mathbf{V}}}(d + nq, \mathbf{D}v_{11}^{*-1} + \textstyle\sum_{i=1}^{n} \mathbf{Y}_i' \mathbf{U}^{-1} \mathbf{Y}_i) I(\mathbf{V}_{11} = 1). \tag{2.7}$$

This component, $HIW(d + nq, \mathbf{D}v_{11}^{*-1} + \sum_{i=1}^{n} \mathbf{Y}_i' \mathbf{U}^{-1} \mathbf{Y}_i) I(\mathbf{V}_{11} = 1)$, is the HIW distribution conditioned on the $1-1$ element of the variance matrix set at unity.

These distributions form the basis of Gibbs sampling for $p(\mathbf{U}, \mathbf{V}, v_{11}^* \mid \mathbf{Y})$. This involves iterative resampling from the hyper-inverse Wishart, inverse gamma and new conditional hyper-inverse Wishart distributions. Simulation of the former is

based on Carvalho et al. (2007), while sampling the latter can be done as follows. From Lemma 2.18 of Lauritzen (1996), I can always find a perfect ordering of the nodes in $G_{\mathbf{V}}$ so that node 1 is in the first clique, say C, and then initialise the hyper-inverse Wishart sampler of Carvalho et al. (2007) to begin with a simulation of the implied conditional inverse Wishart distribution for the covariance matrix on that first clique. Sampling $\mathbf{V}_C$ from an inverse Wishart distribution conditional on the first diagonal element set to unity is straightforward.

2.5 Example: A simulated random sample

A sample of size $n = 48$ was drawn from the $(q = 8) \times (p = 7)$ dimensional $N(\mathbf{0}, \mathbf{U}, \mathbf{V})$ distribution where, using $\cdot$ to denote zeros to highlight structure, the precision matrices are

$$\mathbf{\Lambda} = \begin{pmatrix} 1{\cdot}85 & \text{-}0{\cdot}09 & \text{-}0{\cdot}65 & \cdot & \text{-}0{\cdot}24 & 0{\cdot}45 & \cdot \\ \text{-}0{\cdot}09 & 0{\cdot}21 & 0{\cdot}08 & \cdot & \cdot & 0{\cdot}14 & \text{-}0{\cdot}13 \\ \text{-}0{\cdot}65 & 0{\cdot}08 & 0{\cdot}58 & 0{\cdot}10 & \cdot & \text{-}0{\cdot}30 & \cdot \\ \cdot & \cdot & 0{\cdot}10 & 0{\cdot}48 & \cdot & \text{-}0{\cdot}10 & \cdot \\ \text{-}0{\cdot}24 & \cdot & \cdot & \cdot & 0{\cdot}70 & \text{-}0{\cdot}17 & \cdot \\ 0{\cdot}45 & 0{\cdot}14 & \text{-}0{\cdot}30 & \text{-}0{\cdot}10 & \text{-}0{\cdot}17 & 0{\cdot}61 & \text{-}0{\cdot}36 \\ \cdot & \text{-}0{\cdot}13 & \cdot & \cdot & \cdot & \text{-}0{\cdot}36 & 3{\cdot}72 \end{pmatrix}$$

and

$$\mathbf{\Omega} = \begin{pmatrix} 0{\cdot}99 & \cdot & \cdot & \text{-}0{\cdot}33 & \cdot & 0{\cdot}05 & \cdot & \cdot \\ \cdot & 3{\cdot}65 & 0{\cdot}33 & \cdot & \text{-}0{\cdot}39 & \text{-}0{\cdot}41 & \cdot & \text{-}0{\cdot}03 \\ \cdot & 0{\cdot}33 & 2{\cdot}23 & \cdot & \cdot & \text{-}0{\cdot}38 & \cdot & \cdot \\ \text{-}0{\cdot}33 & \cdot & \cdot & 1{\cdot}65 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \text{-}0{\cdot}39 & \cdot & \cdot & 2{\cdot}91 & \text{-}0{\cdot}30 & \cdot & \cdot \\ 0{\cdot}05 & \text{-}0{\cdot}41 & \text{-}0{\cdot}38 & \cdot & \text{-}0{\cdot}30 & 4{\cdot}71 & \text{-}0{\cdot}13 & \text{-}0{\cdot}40 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \text{-}0{\cdot}13 & 1{\cdot}07 & \text{-}0{\cdot}26 \\ \cdot & \text{-}0{\cdot}03 & \cdot & \cdot & \cdot & \text{-}0{\cdot}40 & \text{-}0{\cdot}26 & 1{\cdot}45 \end{pmatrix}.$$

First consider analysis on the true graphs under priors with $b = d = 3$ and $\mathbf{B} = 5\mathbf{I}_8, \mathbf{D} = 5\mathbf{I}_7$ and simulation sample size 8000 after an initial, discarded burn-in of 2000 iterations. Figure 4.5 presents some trace plots of Monte Carlo samples. Convergence is rapid and apparently fast-mixing in this simulated examples. The

corresponding posterior means of the precision matrices are

$$\hat{\mathbf{\Lambda}} = \begin{pmatrix} 1{\cdot}86 & -0{\cdot}11 & -0{\cdot}68 & \cdot & -0{\cdot}28 & 0{\cdot}44 & \cdot \\ -0{\cdot}11 & 0{\cdot}28 & 0{\cdot}14 & \cdot & \cdot & 0{\cdot}16 & -0{\cdot}21 \\ -0{\cdot}68 & 0{\cdot}14 & 0{\cdot}68 & 0{\cdot}16 & \cdot & -0{\cdot}33 & \cdot \\ \cdot & \cdot & 0{\cdot}16 & 0{\cdot}59 & \cdot & -0{\cdot}15 & \cdot \\ -0{\cdot}28 & \cdot & \cdot & \cdot & 0{\cdot}75 & -0{\cdot}14 & \cdot \\ 0{\cdot}44 & 0{\cdot}16 & -0{\cdot}33 & -0{\cdot}15 & -0{\cdot}14 & 0{\cdot}71 & -0{\cdot}45 \\ \cdot & -0{\cdot}21 & \cdot & \cdot & \cdot & -0{\cdot}45 & 4{\cdot}14 \end{pmatrix}$$

and

$$\hat{\mathbf{\Omega}} = \begin{pmatrix} 0{\cdot}90 & \cdot & \cdot & -0{\cdot}27 & \cdot & -0{\cdot}02 & \cdot & \cdot \\ \cdot & 3{\cdot}23 & 0{\cdot}50 & \cdot & -0{\cdot}35 & -0{\cdot}22 & \cdot & -0{\cdot}12 \\ \cdot & 0{\cdot}50 & 2{\cdot}14 & \cdot & \cdot & -0{\cdot}37 & \cdot & \cdot \\ -0{\cdot}27 & \cdot & \cdot & 1{\cdot}46 & \cdot & \cdot & \cdot & \cdot \\ \cdot & -0{\cdot}35 & \cdot & \cdot & 2{\cdot}88 & -0{\cdot}41 & \cdot & \cdot \\ -0{\cdot}02 & -0{\cdot}22 & -0{\cdot}37 & \cdot & -0{\cdot}41 & 4{\cdot}20 & -0{\cdot}29 & -0{\cdot}08 \\ \cdot & \cdot & \cdot & \cdot & \cdot & -0{\cdot}29 & 0{\cdot}91 & -0{\cdot}26 \\ \cdot & -0{\cdot}12 & \cdot & \cdot & \cdot & -0{\cdot}08 & -0{\cdot}26 & 1{\cdot}58 \end{pmatrix}.$$

2.6 Dynamic matrix-variate graphical models for time series

One motivating interest is models for matrix time series data. Carvalho & West (2007a,b) used graphical structuring for a single covariance matrix in a class of multivariate, exchangeable time series models (Quintana & West, 1987; Quintana, 1992; West & Harrison, 1997) that has been widely used in financial time series (Quintana et al., 2003; Carvalho & West, 2007a,b). Using the theory and methods for matrix normal models developed above, I am now able to extend to matrix time series involving two covariance matrices and associated graphical models. In the notation below, the work of Carvalho & West (2007a,b) is the special case of vector data with $q = 1$, $\mathbf{U}$ fixed, and inference on $(\mathbf{V}, G_{\mathbf{V}})$ only.

A $q \times p$ matrix-variate times series $\mathbf{Y}_t$ follows the dynamic linear model

$$\begin{aligned} \mathbf{Y}_t &= (\mathbf{I}_q \otimes \mathbf{F}_t')\mathbf{\Theta}_t + \boldsymbol{\nu}_t, \quad \boldsymbol{\nu}_t \sim N(\mathbf{0}, \mathbf{U}, \mathbf{V}) \\ \mathbf{\Theta}_t &= (\mathbf{I}_q \otimes \mathbf{G}_t)\mathbf{\Theta}_{t-1} + \mathbf{\Upsilon}_t, \quad \mathbf{\Upsilon}_t \sim N(\mathbf{0}, \mathbf{U} \otimes \mathbf{W}_t, \mathbf{V}) \end{aligned}$$

for $t = 1, 2, \ldots,$ where

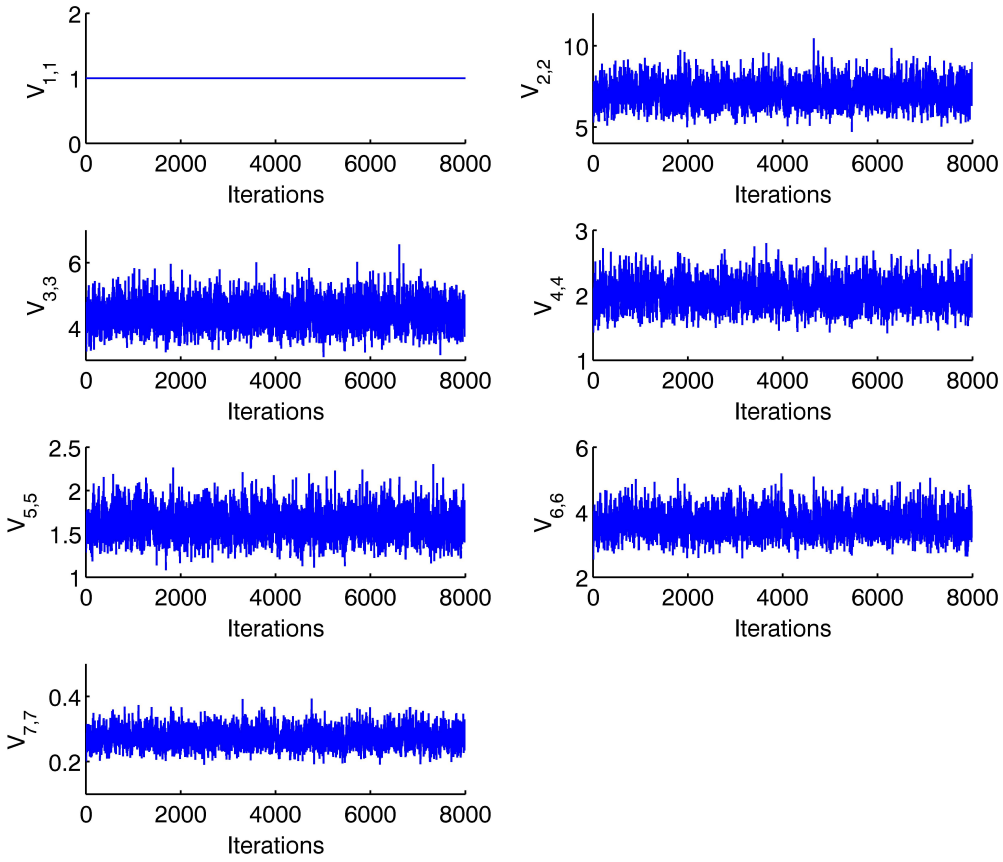

FIGURE 2.1: MCMC traceplots of diagonal elements in $\mathbf{V}$ in the analysis of the simulated random sample of Section 2.5. This illustrates the stability and fast-mixing of the MCMC that is consistent across all parameters in $\mathbf{U}$ and $\mathbf{V}$.

(a) $\mathbf{Y}_t = (\mathbf{Y}_{t,ij})$, the $q \times p$ matrix observation at time t;

(b) $\mathbf{\Theta}_t = (\mathbf{\Theta}_{t,ij})$, the $qs \times p$ state matrix comprised of $q \times p$ state vectors $\mathbf{\Theta}_{t,ij}$ each of dimension $s \times 1$;

(c) $\mathbf{\Upsilon}_t = (\boldsymbol{\omega}_{t,ij})$, the $qs \times p$ matrix of state evolution innovations comprised of $q \times p$ innovation vectors $\boldsymbol{\omega}_{t,ij}$ each of dimension $s \times 1$;

(d) $\boldsymbol{\nu}_t = (nu_{t,ij})$, the $q \times p$ matrix of observational errors;

(e) $\mathbf{W}_t$ is the $s \times s$ innovation covariance matrix at time t;

(f) for all t, the $s-$vector $\mathbf{F}_t$ and $s \times s$ state evolution matrix $\mathbf{G}_t$ are known.

Also, $\boldsymbol{\Upsilon}_t$ follows a matrix-variate normal distribution with mean $\mathbf{0}$, left covariance matrix $\mathbf{U} \otimes \mathbf{W}_t$ and right covariance matrix $\mathbf{V}$. In terms of scalar elements, I have $q \times p$ univariate models with individual $s-$vector state parameters, namely

$$\begin{aligned} \text{Observation:} \quad Y_{t,ij} &= \mathbf{F}_t' \boldsymbol{\Theta}_{t,ij} + \nu_{t,ij}, \quad \nu_{t,ij} \sim N(0, u_{ii} v_{jj}) \\ \text{Evolution:} \quad \boldsymbol{\Theta}_{t,ij} &= \mathbf{G}_t \boldsymbol{\Theta}_{t-1,ij} + \boldsymbol{\omega}_{t,ij}, \quad \boldsymbol{\omega}_{t,ij} \sim N(\mathbf{0}, u_{ii} v_{jj} \mathbf{W}_t) \end{aligned} \tag{2.8}$$

for each i, j and t. Each of the scalar series shares the same $\mathbf{F}_t$ and $\mathbf{G}_t$ elements, and the reference to the model as one of exchangeable time series reflects these symmetries. In the example below $\mathbf{F}_t = \mathbf{F}$ and $\mathbf{G}_t = \mathbf{G}$ as in many practical models, but the model class includes dynamic regressions when $\mathbf{F}_t$ involves predictor variables. This form of model is a standard specification (West & Harrison, 1997) in which the correlation structures induced by $\mathbf{U}$ and $\mathbf{V}$ affect both the observation and evolution errors; for example, if u_{ij} is large and positive, vector series $\mathbf{Y}_{t,\star i}$ and $\mathbf{Y}_{t,\star j}$ will show concordant behavior in movement of their state vectors and in observational variation about their levels. Specification of the entire sequence of $\mathbf{W}_t$ in terms of discount factors (West & Harrison, 1997) is also standard practice, typically using multiple discount factors related to components of the state vector and their expected degrees of random change in time, as illustrated in the example below. The innovations here concern graphical modelling and inference on $(\mathbf{U}, \mathbf{V})$. Key theory, conditional on $\mathbf{U}, \mathbf{V}$, concerns the conjugate sequential learning and forecasting as data is processed, as follows.

Theorem 1. *Define* $D_t = \{D_{t-1}, \mathbf{Y}_t\}$ *for* $t = 1, 2, \ldots,$ *with* D_0 *representing prior information. With initial prior* $(\boldsymbol{\Theta}_0 \mid \mathbf{U}, \mathbf{V}, D_0) \sim N(\mathbf{m}_0, \mathbf{U} \otimes \mathbf{C}_0, \mathbf{V})$ *I have, for all* t :

(a) Posterior at $t-1$: $(\mathbf{\Theta}_{t-1} \mid D_{t-1}, \mathbf{U}, \mathbf{V}) \sim N(\mathbf{m}_{t-1}, \mathbf{U} \otimes \mathbf{C}_{t-1}, \mathbf{V})$

(b) Prior at t : $(\mathbf{\Theta}_t \mid D_{t-1}, \mathbf{U}, \mathbf{V}) \sim N(\mathbf{a}_t, \mathbf{U} \otimes \mathbf{R}_t, \mathbf{V})$ *where* $\mathbf{a}_t = (\mathbf{I}_n \otimes \mathbf{G}_t)\mathbf{m}_{t-1}$ *and* $\mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t' + \mathbf{W}_t$.

(c) One-step forecast at $t-1$: $(\mathbf{Y}_t \mid D_{t-1}, \mathbf{U}, \mathbf{V}) \sim N(\mathbf{f}_t, \mathbf{U}q_t, \mathbf{V})$ *with forecast mean matrix* $\mathbf{f}_t = (\mathbf{I}_n \otimes \mathbf{F}_t'\mathbf{G}_t)\mathbf{m}_{t-1}$ *and scalar* $q_t = \mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t + 1$.

(d) Posterior at t : $(\mathbf{\Theta}_t \mid D_t, \mathbf{U}, \mathbf{V}) \sim N(\mathbf{m}_t, \mathbf{U} \otimes \mathbf{C}_t, \mathbf{V})$ *with* $\mathbf{m}_t = \mathbf{a}_t + (\mathbf{I}_q \otimes \mathbf{A}_t)\mathbf{e}_t$ *and* $\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t\mathbf{A}_t'q_t$ *where* $\mathbf{A}_t = \mathbf{R}_t\mathbf{F}_t/q_t$ *and* $\mathbf{e}_t = \mathbf{Y}_t - \mathbf{f}_t$.

Proof. This results from the theory of multivariate models applied to $\mathrm{vec}(\mathbf{Y}_t)$ (West & Harrison, 1997). The main novelty here concerns the separability of covariance structures. That is: *(a)* for all t, the distributions for state matrices have separable covariance structures; for example, $(\mathbf{\Theta}_t \mid D_t, \mathbf{U}, \mathbf{V})$ is such that $\mathrm{cov}\{\mathrm{vec}(\mathbf{\Theta}_t) \mid \mathrm{D_t}, \mathbf{U}, \mathbf{V}\} = \mathbf{V} \otimes \mathbf{U} \otimes \mathbf{C}_t$; *(b)* the sequential updating equations for the set of $qs \times p$ state matrices are implemented in parallel based on computations for the univariate component models, each of them involving the same scalar q_t, $s-$vector $\mathbf{A}_t$ and $s \times s$ matrices $\mathbf{R}_t, \mathbf{C}_t$ at time t. □

Suppose now that $\mathbf{U}$ and $\mathbf{V}$ are constrained by graphs $G_{\mathbf{U}}$ and $G_{\mathbf{V}}$, with priors as in equation (2.3) and sparsity priors over the graphs. Given data over $t = 1, \ldots, n$, the sequential updating analysis on $(G_{\mathbf{U}}, G_{\mathbf{V}})$ leads to the full joint density

$$p(\mathbf{Y}_1, \ldots, \mathbf{Y}_n \mid \mathbf{U}, \mathbf{V}) = \prod_{t=1}^{n} p(\mathbf{Y}_t \mid \mathbf{U}, \mathbf{V}, D_{t-1}) = \prod_{t=1}^{n} N(\mathbf{e}_t \mid \mathbf{0}, q_t\mathbf{U}, \mathbf{V}),$$

marginalised with respect to all state vectors. The one-step forecast error matrices $\mathbf{e}_t$ are conditionally independent matrix normal variates. Apart from the scalars q_t, this is essentially the framework of Section 2.2. Thus, with a small change to insert the q_t, I am able to directly fit and explore dynamic graphical models using the analysis for random samples with embedded sequential updating computations.

2.7 A macro-economic example

An example concerns exploration of conditional dependence structures in macroeconomic time series related to US labour market employment. The data are Current Employment Statistics for the 8 US states New Jersey, New York, Massachusetts, Georgia, North Carolina, Virginia, Illinois and Ohio. I explore these data across 9 industrial sectors: construction, manufacturing,transportation and utilities, information, financial activities, professional and business services, education and health services, leisure and hospitality, and government. In our model framework, I have $q = 8, p = 9$ and monthly data over several years. Then $\mathbf{U}$ characterises the residual conditional dependencies among states while $\mathbf{V}$ does the same for industrial sectors, in the context of an overall model that incorporates time-varying state parameters for underlying trend and annual seasonal structure in the series. Trend and seasonal elements are represented in standard form, the former as random walks and the latter as randomly varying seasonal effects. Specifically, in month t, the monthly employment change in state i and sector j is $\mathbf{Y}_{t,ij}$, modelled as a first-order polynomial/seasonal effect model (West & Harrison, 1997) with the state vector comprising a local level parameter and 12 seasonal factors, so that the state dimension is $s = 13$.

The univariate models of equation (2.8) have state vectors $\boldsymbol{\Theta}_{t,ij} = (\mu_{t,ij}, \boldsymbol{\phi}_{t,ij})'$ where $\mu_{t,ij}$ is the local level and $\boldsymbol{\phi}_{t,ij} = (\phi_{t,ij,k}, \phi_{t,ij,k+1}, \cdots, \phi_{t,ij,11}, \phi_{t,ij,0}, \cdots, \phi_{t,ij,k-1})$ contains current monthly seasonal factors, subject to $\mathbf{1}'\boldsymbol{\phi}_{t,ij} = 0$ for all i, j and t. Further, $\mathbf{F}_t = \mathbf{F}$ (13×1) and $\mathbf{G}_t = \mathbf{G}$ (13×13) for all t, where $\mathbf{F}' = (1, 1, 0, \cdots, 0)$. The state matrix $\mathbf{G}$ and the sequence of state evolution covariance matrices $\mathbf{W}_t$ (13×13) are

$$\mathbf{G} = \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{P} \end{pmatrix} \text{ with } P = \begin{pmatrix} \mathbf{0} & \mathbf{I}_{11} \\ 1 & \mathbf{0}' \end{pmatrix}, \text{ and } \mathbf{W}_t = \begin{pmatrix} W_{t,\mu} & \mathbf{0}' \\ \mathbf{0} & \mathbf{W}_{t,\phi} \end{pmatrix},$$

with the latter having entries as follows. The univariate $W_{t,\mu}$ and 12×12 matrix

$\mathbf{W}_{t,\phi}$ are defined via discount factors δ_l and δ_s and the corresponding block components of $\mathbf{C}_t$ as $W_{t,\mu} = C_{t-1,\mu}(1-\delta_l)/\delta_l$ and $\mathbf{W}_{t,\phi} = \mathbf{P}\mathbf{C}_{t-1,\phi}\mathbf{P}'(1-\delta_s)/\delta_s$ for each t. The discount factor δ_l reflects the rate at which the levels $\mu_{t,ij}$ are expected to vary between months, with $100(\delta_l^{-1}-1)\%$ of information on these parameters decaying each month. The factor δ_s plays the same role for seasonal parameters. I use $\delta_l = 0{\cdot}9$, $\delta_s = 0{\cdot}95$ to allow more adaptation to level changes than seasonal factors (West & Harrison, 1997); results, in terms of graphical model search and structure, are substantially similar using other values in appropriate ranges. In application, I can estimate discount factors and also extend the model to allow changes in discount factors to model change-points and other events impacting the series, based on monitoring and intervention methods (Pole et al., 1994; West & Harrison, 1997). Such considerations are secondary to our purposes in using this model for illustration of computational model search analysis for $(\mathbf{U}, \mathbf{V}, G_{\mathbf{U}}, G_{\mathbf{V}})$, but practically very germane. Model completion uses initial, vague priors with $\mathbf{m}_0 = \mathbf{0}$, the 104×9 matrix, and $\mathbf{C}_0 = 100\mathbf{I}_{13}$. The constraint that $\mathbf{1}'\boldsymbol{\phi}_{t,ij} = 0$ is imposed by transforming $\mathbf{m}_0$ and $\mathbf{C}_0$ as discussed in West & Harrison (1997).

Applying this model, I aim to detect and estimate sustained movement and changes in trend and seasonality, generating on-line detrended and deseasonalised estimates matrix series $\mathbf{e}_t$ whose row and column covariance patterns are defined by the parameters $\mathbf{U}, \mathbf{V}$. Figure 2.2 through Figure 2.5 display the observed time series and one-step ahead forecasts for all of the eight sections and nine states.

The standardised residuals $\mathbf{e}_t/\sqrt{q_t}$ can now be used as independent and identically-distributed random matrices to draw inference about $(\mathbf{U}, \mathbf{V})$. Priors for $(\mathbf{U}, \mathbf{V})$ use $\mathbf{B} = 5\mathbf{I}_8$, $\mathbf{D} = 5\mathbf{I}_9$ and $b = d = 3$, reflecting the range of residual variation. Assume given full graphs for both $G_{\mathbf{U}}$ and $G_{\mathbf{V}}$. Table 2.1 and 2.2 respectively display the Monte Carlo posterior estimates of two precision matrices and two correlation matrices using the sample mean of the 5000 simulated precision and correlation matrices.

Table 2.1: Posterior means of precision matrices in analysis of the matrix econometric time series data under full graphs. The 8 US states are: NJ, New Jersey; NY, New York; MA, Massachusetts; GA, Georgia; NC, North Carolina; VA, Virginia; IL, Illinois; OH, Ohio. The 9 industrial sectors are: C, industrial construction; M, manufacturing; T&U, transportation & utilities; I, information; FA, financial activities; P&BS, professional & business; E&H, services, education & health; L&H, services, leisure & hospitality; G, government.

	NJ	NY	MA	GA	NC	VA	IL	OH
NJ	0.86	−0.02	−0.16	−0.04	−0.01	−0.14	−0.06	−0.09
NY		0.25	−0.20	−0.02	0.01	−0.10	−0.01	−0.02
MA			1.76	−0.05	−0.11	−0.20	−0.08	−0.06
GA				0.75	−0.07	−0.04	−0.04	−0.06
NC					1.09	−0.18	−0.03	−0.11
VA						1.32	−0.03	−0.07
IL							0.43	−0.08
OH								0.64

	C	M	T&U	I	FA	P&BS	E&H	L&H	G
C	1.17	0.00	−0.09	−0.06	−0.10	−0.09	−0.08	−0.19	−0.02
M		0.43	−0.06	−0.05	−0.00	−0.04	−0.04	−0.01	−0.00
T&U			0.41	0.03	−0.10	−0.04	−0.04	−0.08	0.00
I				1.81	−0.04	−0.04	0.03	0.01	−0.04
FA					1.74	−0.08	−0.02	0.01	−0.01
P&BS						0.38	−0.01	−0.11	0.00
E&H							0.75	−0.05	−0.01
L&H								0.77	−0.00
G									0.22

These tables suggest that there might be many elements that are close to zero in precision matrices. Assuming full graphs for $(\mathbf{U}, \mathbf{V})$ might cause the ignorance of these sparse structures. I discuss how to conduct the fully Bayesian analysis to address the graphical model uncertainty and model determination in Chapter 3.

Table 2.2: Posterior means of correlation matrices in analysis of the matrix econometric time series data under full graphs. The 8 US states are: NJ, New Jersey; NY, New York; MA, Massachusetts; GA, Georgia; NC, North Carolina; VA, Virginia; IL, Illinois; OH, Ohio. The 9 industrial sectors are: C, industrial construction; M, manufacturing; T&U, transportation & utilities; I, information; FA, financial activities; P&BS, professional & business; E&H, services, education & health; L&H, services, leisure & hospitality; G, government.

	NJ	NY	MA	GA	NC	VA	IL	OH
NJ	1	0.17	0.24	0.12	0.11	0.23	0.19	0.21
NY		1	0.38	0.13	0.10	0.28	0.12	0.17
MA			1	0.14	0.17	0.28	0.19	0.20
GA				1	0.14	0.14	0.14	0.16
NC					1	0.22	0.13	0.21
VA						1	0.15	0.20
IL							1	0.23
OH								1

	C	M	T&U	I	FA	P&BS	E&H	L&H	G
C	1	0.07	0.22	0.05	0.12	0.24	0.14	0.28	0.04
M		1	0.18	0.06	0.05	0.14	0.09	0.09	0.02
T&U			1	−0.00	0.16	0.21	0.13	0.24	0.01
I				1	0.03	0.06	−0.02	0.01	0.07
FA					1	0.15	0.05	0.07	0.02
P&BS						1	0.09	0.28	0.02
E&H							1	0.13	0.03
L&H								1	0.02
G									1

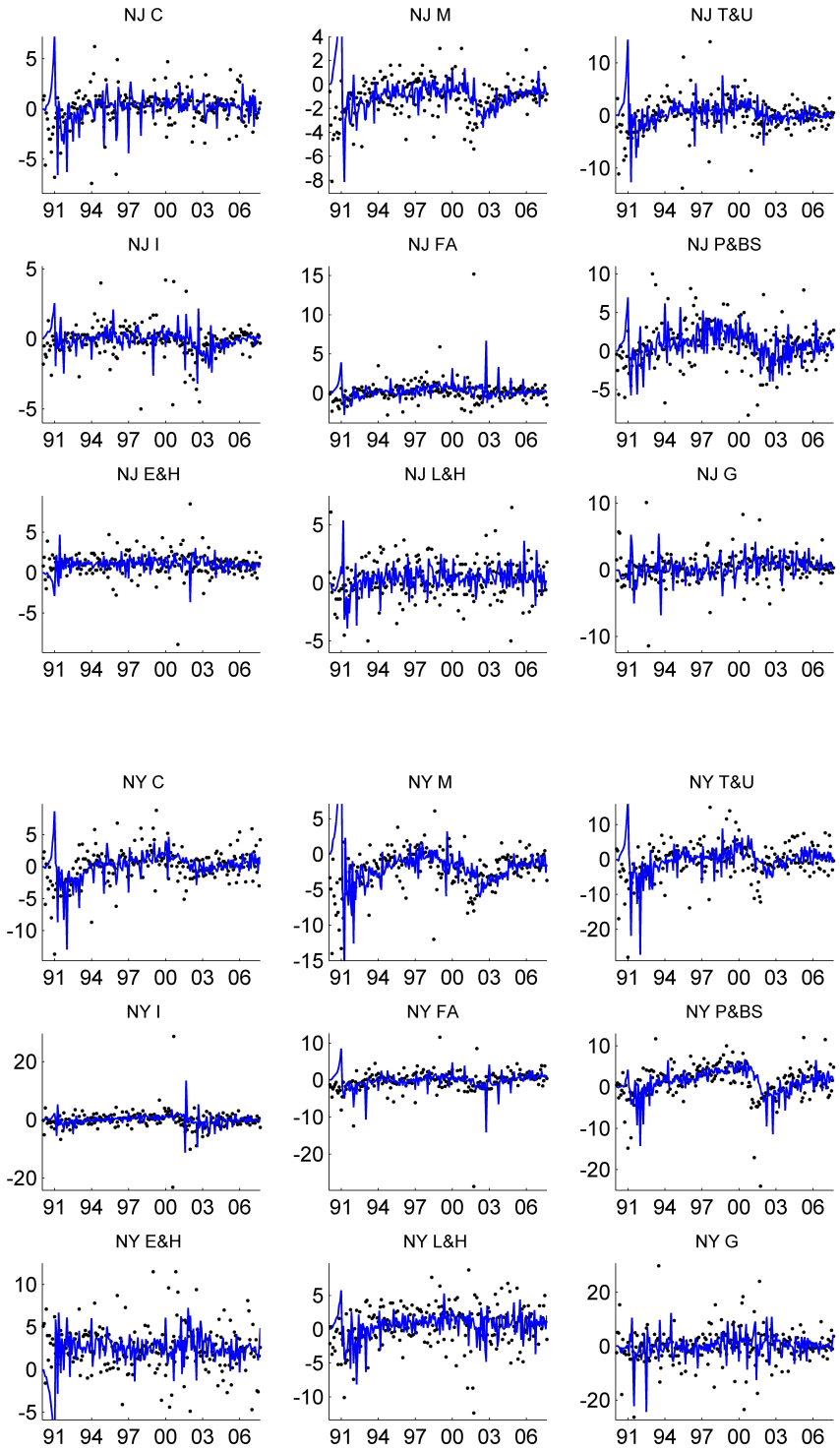

FIGURE 2.2: Time series in the econometric example, plotted over 1990-2007. Monthly changes in employment across nine sectors for NJ and NY together with the one-step ahead forecasts.

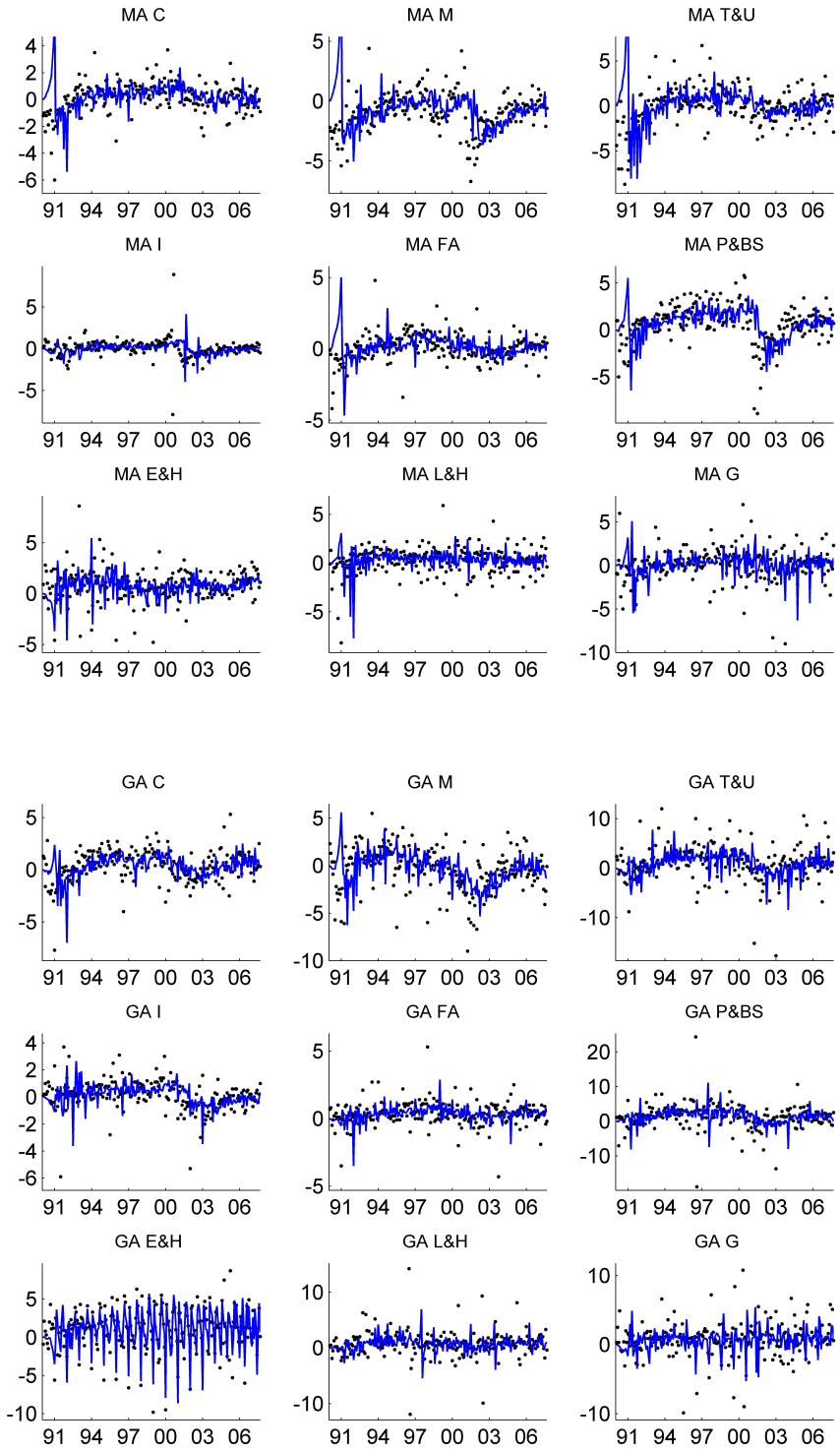

FIGURE 2.3: Time series in the econometric example, plotted over 1990-2007. Monthly changes in employment across nine sectors for MA and GA together with the one-step ahead forecasts.

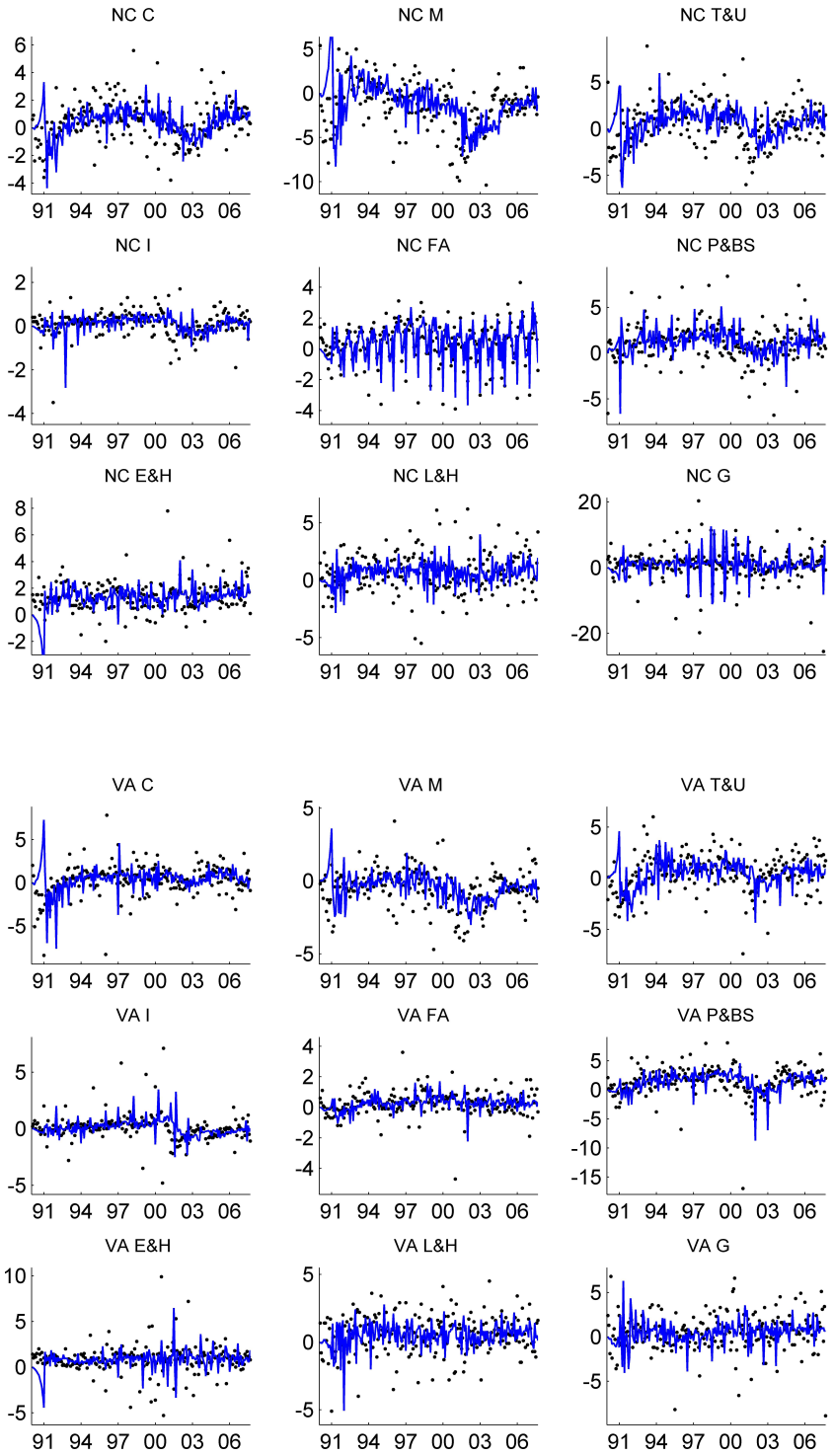

FIGURE 2.4: Time series in the econometric example, plotted over 1990-2007. Monthly changes in employment across nine sectors for NC and VA together with the one-step ahead forecasts.

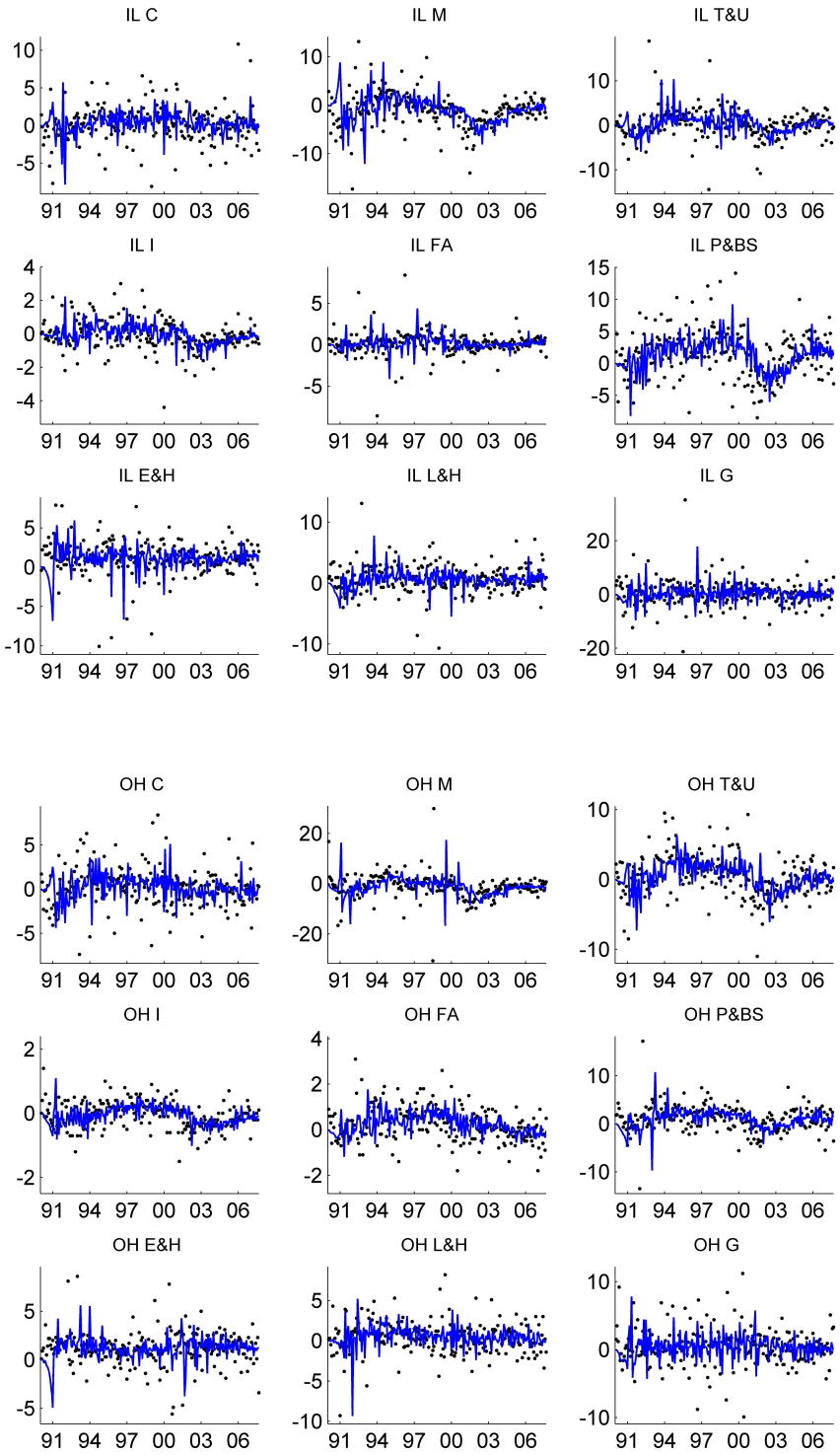

FIGURE 2.5: Time series in the econometric example, plotted over 1990-2007. Monthly changes in employment across nine sectors for IL and OH together with the one-step ahead forecasts.

3

Matrix normal graphical model determination

3.1 Marginal likelihood

Exploration of uncertainty about graphical model structures involves consideration of the marginal likelihood function over graphs. For any pair $(G_{\mathbf{U}}, G_{\mathbf{V}})$, this is

$$p(\mathbf{Y}) \equiv p(\mathbf{Y} \mid G_{\mathbf{U}}, G_{\mathbf{V}}) \;\; = \;\; \int p(\mathbf{Y} \mid \mathbf{U}, \mathbf{V}) p(\mathbf{U}) p(\mathbf{V}) \mathrm{d}\mathbf{U}\, \mathrm{d}\mathbf{V}.$$

The priors in the integrand depend on the graphs although I drop that in the notation for clarity. In multivariate models, marginal likelihoods can be evaluated in closed form on decomposable graphs (Giudici, 1996; Giudici & Green, 1999; Jones et al., 2005; Carvalho & West, 2007a,b). In our matrix models, the integral cannot be evaluated but I can generate useful approximations via use of Candidate's formula (Besag, 1989; Chib, 1995). Write $\mathbf{\Theta} = \{\mathbf{U}, \mathbf{V}, v_{11}^*\}$ for all parameters, and suppose that I can evaluate $p(\boldsymbol{\theta} \mid \mathbf{Y})$ for some subset of parameters $\boldsymbol{\theta} \in \mathbf{\Theta}$; Candidate's formula gives the marginal likelihood via the identity $p(\mathbf{Y}) = p(\mathbf{Y} \mid \boldsymbol{\theta})/p(\boldsymbol{\theta} \mid \mathbf{Y})$. Applying this requires that I estimate components of the numerator or denominator. Choosing $\boldsymbol{\theta}$ to maximally exploit analytic integration is key, and different choices

that integrate over different subsets of parameters will lead to different, parallel approximations of $p(\mathbf{Y})$ that can be compared. I use two approximations based on marginalisation over desirably disjoint parameter subsets, namely

(A) $p(\mathbf{Y}) = p(\mathbf{Y}, v_{11}^*, \mathbf{U})/p(v_{11}^*, \mathbf{U} \mid \mathbf{Y})$ at any chosen value of $\boldsymbol{\theta} = \{v_{11}^*, \mathbf{U}\}$, and

(B) $p(\mathbf{Y}) = p(\mathbf{Y}, \mathbf{V})/p(\mathbf{V} \mid \mathbf{Y})$ at any value of $\boldsymbol{\theta} = \mathbf{V}$.

I estimate the components of these equations that have no closed form, then plug-in chosen values $\mathbf{U}, \mathbf{V}, v_{11}^*$, such as approximate posterior means, to provide two estimates of $p(\mathbf{Y})$.

For (A), first rewrite as

$$p(\mathbf{Y}) = \frac{p(\mathbf{Y}, v_{11}^*, \mathbf{U})p(\mathbf{V} \mid v_{11}^*, \mathbf{U}, \mathbf{Y})}{p(v_{11}^*, \mathbf{U} \mid \mathbf{Y})p(\mathbf{V} \mid v_{11}^*, \mathbf{U}, \mathbf{Y})} = \frac{p(\mathbf{Y} \mid \mathbf{V}, v_{11}^*, \mathbf{U})p(\mathbf{U})p(\mathbf{V} \mid v_{11}^*)p(v_{11}^*)}{p(v_{11}^*, \mathbf{U} \mid \mathbf{Y})p(\mathbf{V} \mid v_{11}^*, \mathbf{U}, \mathbf{Y})}.$$

The numerator terms are each easily computed at any $\{\mathbf{V}, v_{11}^*, \mathbf{U}\}$. The second denominator term $p(\mathbf{V} \mid v_{11}^*, \mathbf{U}, \mathbf{Y})$ has an easily evaluated closed form, as in the Gibbs sampling step. The first denominator term may be approximated by

$$\begin{aligned} p(v_{11}^*, \mathbf{U} \mid \mathbf{Y}) &= \int p(v_{11}^* \mid \mathbf{Y}, \mathbf{V})p(\mathbf{U} \mid \mathbf{Y}, \mathbf{V}, v_{11}^*)p(\mathbf{V} \mid \mathbf{Y})d\mathbf{V} \\ &\approx \frac{1}{M}\sum_{j=1}^{M} p(v_{11}^* \mid \mathbf{Y}, \mathbf{V}_j)p(\mathbf{U} \mid \mathbf{Y}, \mathbf{V}_j, v_{11}^*) \end{aligned}$$

where the sum is over posterior draws $\mathbf{V}_j$; this is easy to compute as it is a sum of the product of inverse gamma and hyper-inverse Wishart densities.

For (B), the numerator can be analytically evaluated as

$$
\begin{aligned}
p(\mathbf{V},\mathbf{Y}) &= \int p(\mathbf{Y},\mathbf{U},\mathbf{V},v_{11}^*)\mathrm{d}\mathbf{U}\,\mathrm{d}v_{11}^* \\
&= \int N(\mathbf{Y}_{1:n}\mid\mathbf{0},\mathbf{U},\mathbf{V})HIW_{G_\mathbf{U}}(\mathbf{U}\mid b,\mathbf{B})HIW(v_{11}^*\mathbf{V}|d,\mathbf{D})(v_{11}^*)^{\nu-1}\mathrm{d}\mathbf{U}\,\mathrm{d}v_{11}^* \\
&= \int N(\mathbf{Y}_{1:n}\mid\mathbf{0},\mathbf{U},\mathbf{V})HIW_{G_\mathbf{U}}(\mathbf{U}\mid b,\mathbf{B})\mathrm{d}\mathbf{U}\times \\
&\quad\times\int HIW(v_{11}^*\mathbf{V}|d,\mathbf{D})(v_{11}^*)^{\nu-1}\mathrm{d}v_{11}^* \\
&= \frac{(2\pi)^{-\frac{nqp}{2}}|\mathbf{V}|^{-\frac{nq}{2}}H(b,\mathbf{B},G_\mathbf{U})}{H(b+np,\mathbf{B}+\sum_i^n\mathbf{Y}_i\mathbf{V}^{-1}\mathbf{Y}_i',G_\mathbf{U})}\times \\
&\quad\times\frac{\prod_{P_\mathbf{V}\in\mathcal{P}_\mathbf{V}}|\mathbf{V}_{P_\mathbf{V}}|^{-\frac{d+2|P_\mathbf{V}|}{2}}H(d,\mathbf{D},G_\mathbf{V})}{\prod_{S_\mathbf{V}\in\mathcal{S}_\mathbf{V}}|\mathbf{V}_{S_\mathbf{V}}|^{-\frac{d+2|S_\mathbf{V}|}{2}}H(c,\mathrm{tr}(\mathbf{D}\mathbf{V}^{-1}),1)} \\
&= \frac{q_\mathbf{V}(2\pi)^{-nqp/2}H(b,\mathbf{B},G_\mathbf{U})H(d,\mathbf{D},G_\mathbf{V})}{H(b+nq,\mathbf{B}+\sum_i^n\mathbf{Y}_i\mathbf{V}^{-1}\mathbf{Y}_i,G_\mathbf{U})H(a,\mathrm{tr}(\mathbf{D}\mathbf{V}^{-1}),1)}
\end{aligned}
$$

where

$$
q_\mathbf{V}=\prod_{P_\mathbf{V}\in\mathcal{P}_\mathbf{V}}|\mathbf{V}_{P_\mathbf{V}}|^{-(nq+d+2|P_\mathbf{V}|)/2}\Big/\prod_{S_\mathbf{V}\in\mathcal{S}_\mathbf{V}}|\mathbf{V}_{S_\mathbf{V}}|^{-(nq+d+2|S_\mathbf{V}|)/2},
$$

the $H(\cdot,\cdot,G_\cdot)$ terms are normalising constants of the corresponding hyper-inverse Wishart distributions (Giudici & Green, 1999; Jones et al., 2005) and

$$
c=\sum_{P_\mathbf{V}\in\mathcal{P}_\mathbf{V}}|P_\mathbf{V}|(2|P_\mathbf{V}|+d)-\sum_{S_\mathbf{V}\in\mathcal{S}_\mathbf{V}}|S_\mathbf{V}|(2|S_\mathbf{V}|+d)-2\nu.
$$

The density function in the denominator is approximated as

$$
p(\mathbf{V}\mid\mathbf{Y})=\int p(\mathbf{V}\mid v_{11}^*,\mathbf{U},\mathbf{Y})p(v_{11}^*,\mathbf{U}\mid\mathbf{Y})\mathrm{d}v_{11}^*\,\mathrm{d}\mathbf{U}\approx\frac{1}{M}\sum_{j=1}^M P(\mathbf{V}\mid\mathbf{Y},\mathbf{U}_j,v_{11,j}^*)
$$

where the sum over posterior draws $(\mathbf{U}_j,v_{11,j}^*)$ can be easily performed, with terms given by conditional hyper-inverse Wishart density evaluations.

3.2 Example: Markov random fields from matrix graphical models

A rather interesting class of matrix graphical structures arises under autoregressive (AR) correlation specifications for the two precision matrices. This generates a novel class of Markov random field models that is of potential interest in application areas such as texture image modelling. I use this construction here for a second, much higher-dimensional synthetic example.

Take $\mathbf{U}$ and $\mathbf{V}$ as covariances matrices of stationary AR process. For example here, I choose $q = p = 60$ taking $\mathbf{U}$ as the 60×60 variance matrix of an AR(5) model with AR parameters $(0.91, -0.44, 0.38, -0.31, 0.22)$ and marginal variance 0.55, and $\mathbf{V}$ as the 60×60 variance matrix of an AR(4) model with AR parameters $(0.47, 0.23, 0.14, -0.19)$ and marginal variance 0.41. This model is used to repeatedly simulate 50 observations and each draw from the model is a sampled Markov random field; the columns of each sample are correlated realisations from the underlying AR(5) model, and the rows correlated realisations of the AR(4) model. Figure 3.1 images the two underlying precision matrices along with two representative samples.

To illustrate model fitting and evaluation, I use a prior specified with $d = b = 3, \mathbf{D} = (d+2)\mathbf{I}_{60}$ and $\mathbf{B} = 0.01(b+2)\mathbf{I}_{60}$. MCMC analysis uses burn-in of 1000 and then saved 2000 samples starting with initial value $\mathbf{V} = \mathbf{I}_{60}$. The MCMC was run repeatedly across a range of models differing in the order of the underlying AR models for rows and columns, exploring all combinations of AR(1) to AR(9) structures for each of the precision matrices. Applying the model marginal likelihood approximation to each model allows us to evaluate model orders. Table 3.1 shows the top 5 models selected by the largest log-marginal likelihood. As can be seen, the true model orders lead to the largest marginal likelihood and, more importantly in terms of assessing the effectiveness of the methodology, the two parallel marginal likelihood assessments are in concordance and differ negligibly on the scale of interest.

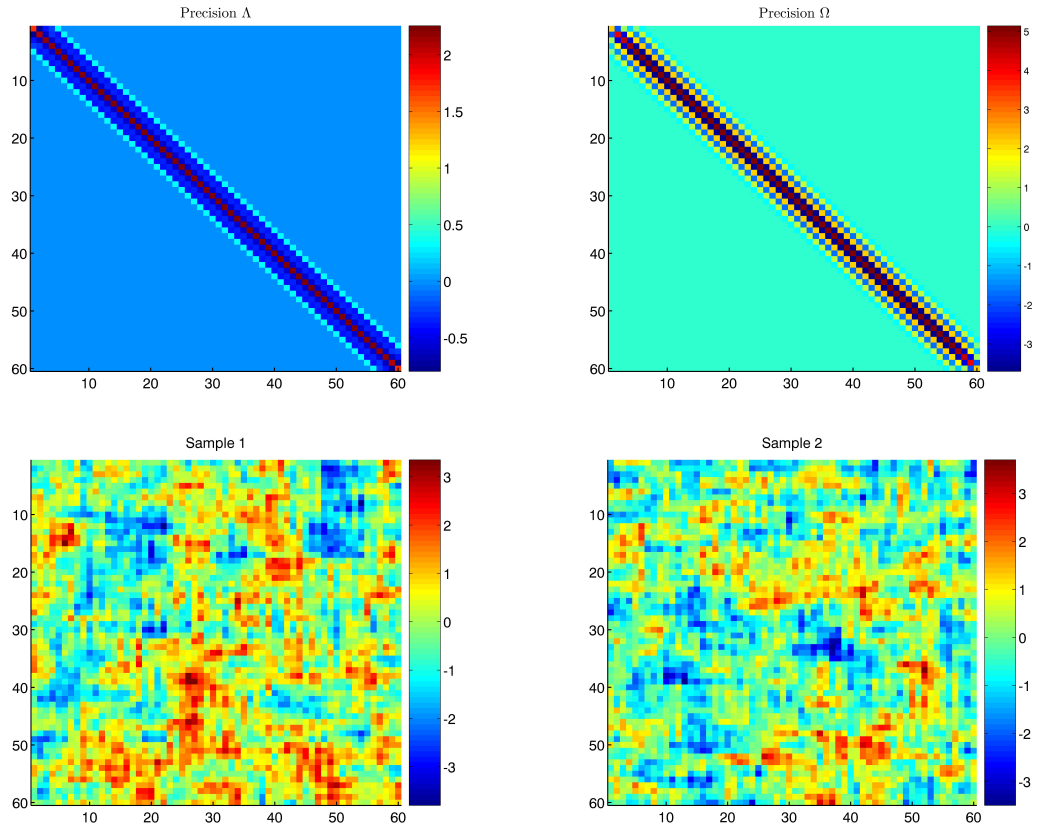

FIGURE 3.1: Images displaying the band structure of the two precision matrices (upper row) used in the MRF 60×60 matrix graphical model example of Section 3.2, together with images of two simulated draws (lower row) from the model.

3.3 Graphical model uncertainty and search

Now admit uncertainty about graphs $(G_{\mathbf{U}}, G_{\mathbf{V}})$ using sparsity-encouraging priors in which edge inclusion indicators are independent Bernoulli variates (Dobra et al., 2004; Jones et al., 2005). Section 1.2.3 has discussed the choice of the edge inclusion probability. I use the default choice of $2/(p-1)$ for a graph with p nodes in all examples in this chapter. I now extend Markov chain Monte Carlo simulation for multivariate graphical models (Giudici & Green, 1999; Jones et al., 2005) to learning on $(G_{\mathbf{U}}, G_{\mathbf{V}})$ in the above matrix model analysis. Our analysis generates multiple graphs with values of approximate posterior probabilities, using the Markov chain

Table 3.1: Relative log-marginal likelihood of the top five models in the MRF graphical model. Each entry is the estimated log-marginal likelihood relative to that of the most likely model on Candidate's method (A).

Graph Structure		log likelihood (A)	log likelihood (B)
V	**U**		
AR(4)	AR(5)	0	0.004
AR(5)	AR(5)	−127.8	−127.9
AR(6)	AR(5)	−234.3	−234.2
AR(4)	AR(6)	−355.8	−355.8

simulation for model search. This relies on the computation of the unnormalised posterior over graphs, $p(G_{\mathbf{U}}, G_{\mathbf{V}} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid G_{\mathbf{U}}, G_{\mathbf{V}})p(G_{\mathbf{U}}, G_{\mathbf{V}})$ involving the marginal likelihood value for any specified model $(G_{\mathbf{U}}, G_{\mathbf{V}})$ at each search step. For the latter, I average the approximate marginal likelihood values from methods (A) and (B). Jones et al. (2005) discuss performance of various stochastic search methods in single multivariate graphical models; for modest dimensions, they recommend simple local-move Metropolis-Hastings steps. Here, given a current pair $(G_{\mathbf{U}}, G_{\mathbf{V}})$, I can apply local moves in $G_{\mathbf{U}}$ space based on the conditional posterior $p(G_{\mathbf{U}} \mid \mathbf{Y}, G_{\mathbf{V}})$, and vice-versa. A candidate $G'_{\mathbf{U}}$ is sampled from a proposal distribution $q(G'_{\mathbf{U}}; G_{\mathbf{U}})$ and accepted with probability

$$\alpha = \min\{\ 1,\ p(G'_{\mathbf{U}} \mid \mathbf{Y}, G_{\mathbf{V}})q(G_{\mathbf{U}}; G'_{\mathbf{U}})/p(G_{\mathbf{U}} \mid \mathbf{Y}, G_{\mathbf{V}})q(G'_{\mathbf{U}}; G_{\mathbf{U}})\ \};$$

our examples use the simple random add/delete edge move proposal of Jones et al. (2005). I then couple this with a similar step using $p(G_{\mathbf{V}} \mid \mathbf{Y}, G_{\mathbf{U}})$ at each iteration. This requires a Markov chain analysis on each graph pair visited in order to evaluate marginal likelihood, so implying a substantial computational burden.

3.4 Examples

3.4.1 Example: A simulated random sample (continued)

I illustrate the marginal likelihood approximation, model uncertainty and search with the simulated random sample example introduced in Section 2.5.

Using the same priors for $\mathbf{U}$ and $\mathbf{V}$ as in Section 2.5, I evaluated the two marginal likelihood estimates under the true graphs at differing simulated sample size ranging from 2000 to 8000. Figure 5.1 gives an implementation check on the concordance of the two marginal likelihood estimates. These are very close and differ negligibly on the log probability scale even at small Monte Carlo sample sizes.

Consider graphical model uncertainty with prior edge inclusion probabilities $2/(q-1)$ for $G_{\mathbf{U}}$ and $2/(p-1)$ for $G_{\mathbf{V}}$. Repeat explorations suggest stability of the marginal likelihood estimation using smaller Monte Carlo sample sizes, and I use 2000 draws within each step of the model search. The add-delete Metropolis-within-Gibbs was run for 20000 iterates starting from empty graphs. Results are essentially replicated starting at the full graphs. The most probable graphs visited, $(\hat{G}_{\mathbf{U}}, \hat{G}_{\mathbf{V}})$, are pictured in Figure 3.3; these are local modes and also have greater posterior probability than the true graphs also displayed, and this model was first visited after 2614 Markov chain steps. The edges in $(\hat{G}_{\mathbf{U}}, \hat{G}_{\mathbf{V}})$ generally have higher posterior edge inclusion probability than those not included; the lowest probability included edge has probability 0·52, while the highest probability excluded edge has probability 0·59. Thus, graphs discovered by highest posterior probability and by aggregating high probability edges are not dramatically different. The modal $\hat{G}_{\mathbf{U}}$ is sparser than the true $G_{\mathbf{U}}$, reflecting the difficulties in identifying very weak signals; for example, the modal graph lacks an edge corresponding to the true $\mathbf{\Omega}_{1,6} = 0.05$, and the posterior probability of that edge is naturally low. One measure of inferred sparsity is the posterior mean of the proportion of edges in each graph; these are about 28%,

59·6% for $G_{\mathbf{U}}, G_{\mathbf{V}}$, respectively. Additional posterior summaries and exploration of the posterior samples suggest clean convergence of the simulation analysis and the Metropolis-Hastings steps over graphs had good empirical acceptance rates of about 26%, 9% for $G_{\mathbf{U}}, G_{\mathbf{V}}$, respectively.

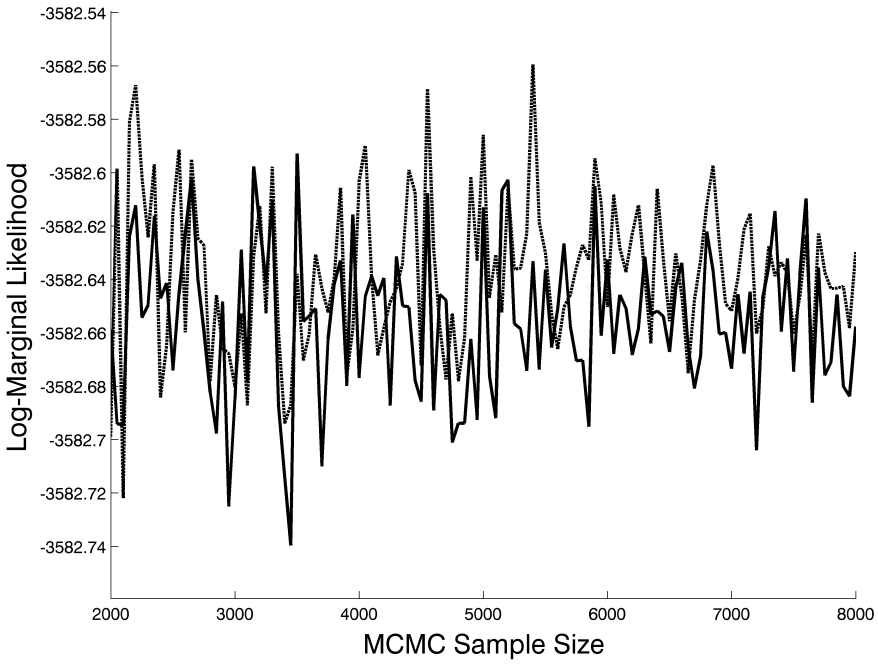

FIGURE 3.2: Log-marginal likelihood values in the simulation example of Section 3.4.1. The two estimates of Section 3.1 were successively re-evaluated and plotted here at differing simulation sample sizes.

3.4.2 Example: A macro-economic example (continued)

I emphasise the practical importance of structured graphical modelling with another example in which the macro-economic data of Section 2.7 are studied.

For the covariance matrices of the residuals, I used the same priors as in Section 2.7. For the priors over graphical model spaces, I used the sparsity-encouraging priors with prior edge inclusion probabilities $2/(q-1)$ for $G_{\mathbf{U}}$ and $2/(p-1)$ for $G_{\mathbf{V}}$. To conduct a fully Bayesian model determination procedure of Section 3.1 and

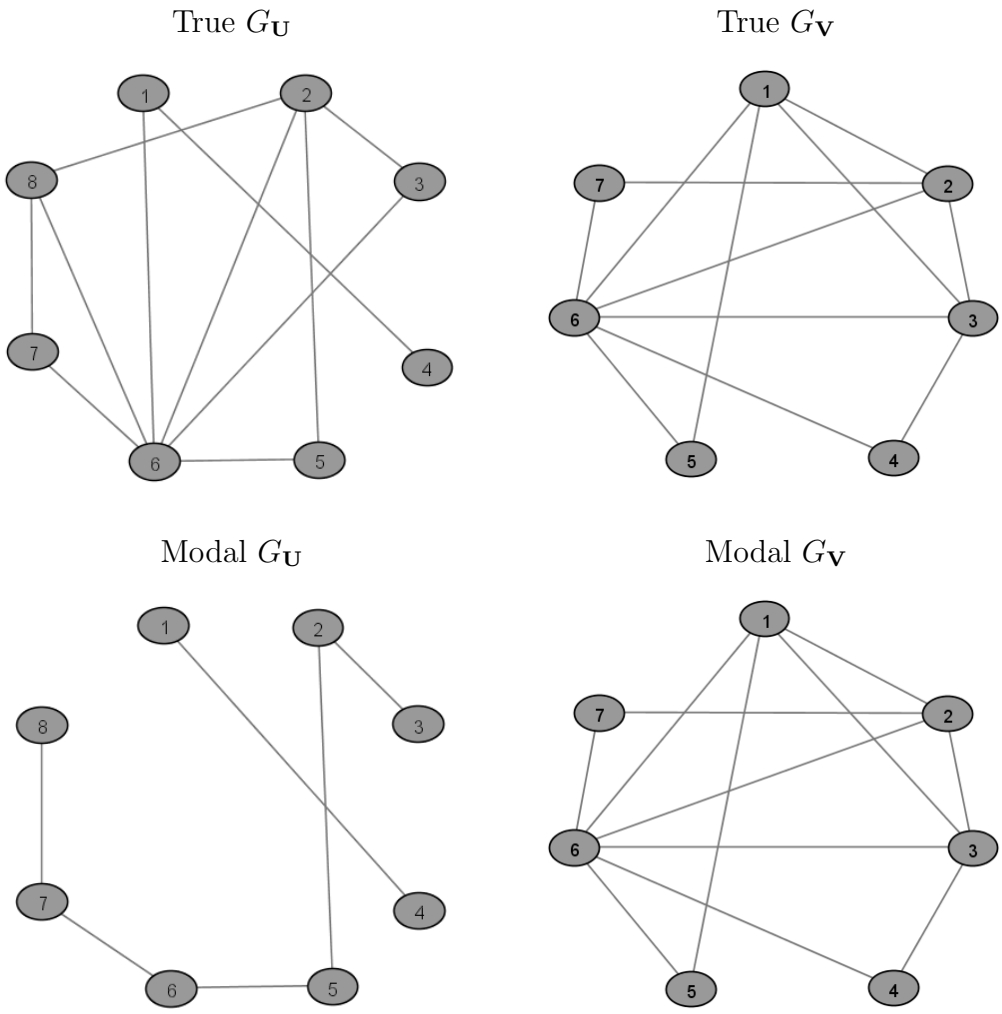

FIGURE 3.3: True graphs in the simulated data example together with graphs of highest posterior probability identified from the analysis.

3.3, I ran the add-delete Metropolis-within-Gibbs sampler for 20000 steps. Two chains were run: one starting at empty graphs and one at full graphs. In Table 3.2, I see that the most probable graph combination, those identified with highest posterior probability and pictured in Figure 3.4, and the acceptance rate in graph spaces $(G_{\mathbf{U}}, G_{\mathbf{V}})$ were insensitive to the starting points. Posterior edge inclusion probabilities are also generally consistent between the two runs; see Table 3.3. In terms of posterior probability and sparsity as measured by the proportion of edges

Table 3.2: Summary of two Metropolis-Hastings chains in graphical model analysis of the econometric time series data. *First row*: analysis initialised at empty graphs. *Second row*: analysis initialised at full graphs. The second column reports the number of Markov chain steps to reach the highest posterior graphs $(G_{\mathbf{U}}, G_{\mathbf{V}})$ found, and the third column gives posterior means of proportions of edges in each graph.

Max log posterior	Graphs to first top graphs visited	Sparsity $(G_{\mathbf{U}}, G_{\mathbf{V}})$	Acceptance rate $(\alpha_{G_{\mathbf{U}}}, \alpha_{G_{\mathbf{V}}})$
−27695·40	401	(72·4%,42·1%)	(7·3%,11·9%)
−27695·43	2194	(73·7%,41·9%)	(7·7%,12·2%)

in a graph, the most probable graphs sit in a region of graph space population by graphs of similar sparsity and posterior probability; see Figure 3.5. The posterior is dense around this mode.

Graphs with high probability in the region of the mode seem to reflect relevant dependencies in the econometric context. There are strongly evident conditional independencies particularly among subsets of the industrial sectors; see Table 3.3. Further, the posterior indicates overall sparsity levels via posterior means of about 73% for the proportion of included edges in the $G_{\mathbf{U}}$ graphs and 42% in the $G_{\mathbf{V}}$ graph. Figure 3.5 further illustrates aspects of the posterior over sparsity for $G_{\mathbf{V}}$.

3.5 Further comments

In Chapter 2 and 3, I have introduced Bayesian analysis of matrix-variate graphical models in random sampling and time series contexts. The main innovations include new priors for matrix normal graphical models, use of the parameter expansion approach, inference via Markov chain Monte Carlo for a specific graphical model, evaluation of marginal likelihoods over graphs using coupled Candidate's formula approximations, and the extension of graphical modelling to matrix time series analysis.

On the use of parameter expansion, Roy & Hobert (2007) and Hobert & Marchev

Table 3.3: Posterior edge inclusion probabilities in graphical model analysis of the matrix econometric time series data. The 8 US states are: NJ, New Jersey; NY, New York; MA, Massachusetts; GA, Georgia; NC, North Carolina; VA, Virginia; IL, Illinois; OH, Ohio. The 9 industrial sectors are: C, industrial construction; M, manufacturing; T&U, transportation & utilities; I, information; FA, financial activities; P&BS, professional & business; E&H, services, education & health; L&H, services, leisure & hospitality; G, government.

	NJ	NY	MA	GA	NC	VA	IL	OH
NJ	1	0·05	1·00	0·55	0·01	1·00	1·00	1·00
NY		1	1·00	0·19	0·00	1·00	0·00	0·59
MA			1	0·98	0·96	1·00	1·00	1·00
GA				1	0·93	0·89	0·75	1·00
NC					1	1·00	0·06	1·00
VA						1	0·31	1·00
IL							1	1·00
OH								1

	C	M	T&U	I	FA	P&BS	E&H	L&H	G
C	1	0·02	1·00	0·16	0·75	1·00	0·99	1·00	0·06
M		1	1·00	0·28	0·02	0·98	0·01	0·03	0·01
T&U			1	0·02	1·00	1·00	0·93	1·00	0·02
I				1	0·06	0·34	0·02	0·01	0·55
FA					1	1·00	0·00	0·04	0·02
P&BS						1	0·02	1·00	0·00
E&H							1	0·75	0·02
L&H								1	0·01
G									1

(2008) provide theoretical support for the method in Gibbs samplers; in our models, this approach induces tractable and computationally accessible posteriors, leads to good mixing of Markov chain simulations, and is theoretically fundamental to the new model/prior framework in addressing identification issues directly and naturally.

On model identification, an alternative approach might use unconstrained hyper-inverse Wishart priors for each of $(\mathbf{U}, \mathbf{V})$ and run the Markov chain Monte Carlo simulation on the unconstrained parameters, similar to a strategy sometimes used in multinomial probit models (McCulloch et al., 2000). It can be argued that this

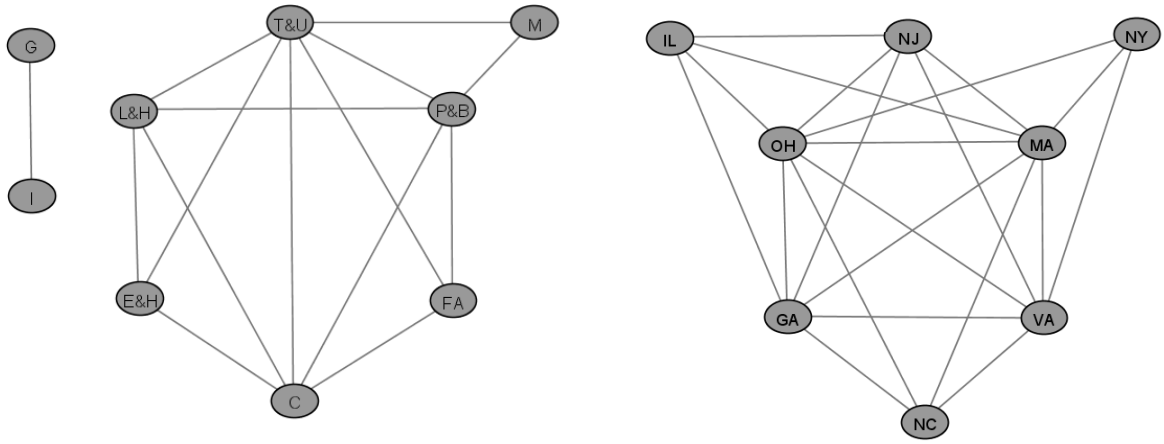

FIGURE 3.4: Highest posterior probability graphs that illustrate aspects of inferred conditional dependencies among industrial sectors and among states in analysis of the econometric time series data.

is computationally less demanding than using our explicitly constrained prior and that inferences can be constructed from the simulation output by transforming to constraint-compatible parameters $(\mathbf{U}v_{11}, \mathbf{V}/v_{11})$. I had considered this, and note that posterior simulation analysis is marginally faster than under the explicitly identified model; in empirical studies, however, I find the computational benefit to be of negligible practical significance. Importantly, this approach relies on a proper prior for the effectively free, unidentified parameter v_{11}, and is sensitive to that choice. More importantly, the implied prior on $(\mathbf{U}v_{11}, \mathbf{V}/v_{11})$ is non-standard and difficult to interpret, and raises questions in prior elicitation and specification; for example, the implied margins for variances are those of ratios of inverse gamma variates and difficult to assess compared to the traditional inverse gamma, and there are now dependencies in priors on left and right covariance matrices. Perhaps most important are resulting effects on approximate marginal likelihoods; in examples I have studied, the approach yields very different marginal likelihoods and the impact of the marginal prior on the unidentified v_{11} plays a key role in that. In contrast, and though very slightly more computationally demanding, the direct and explic-

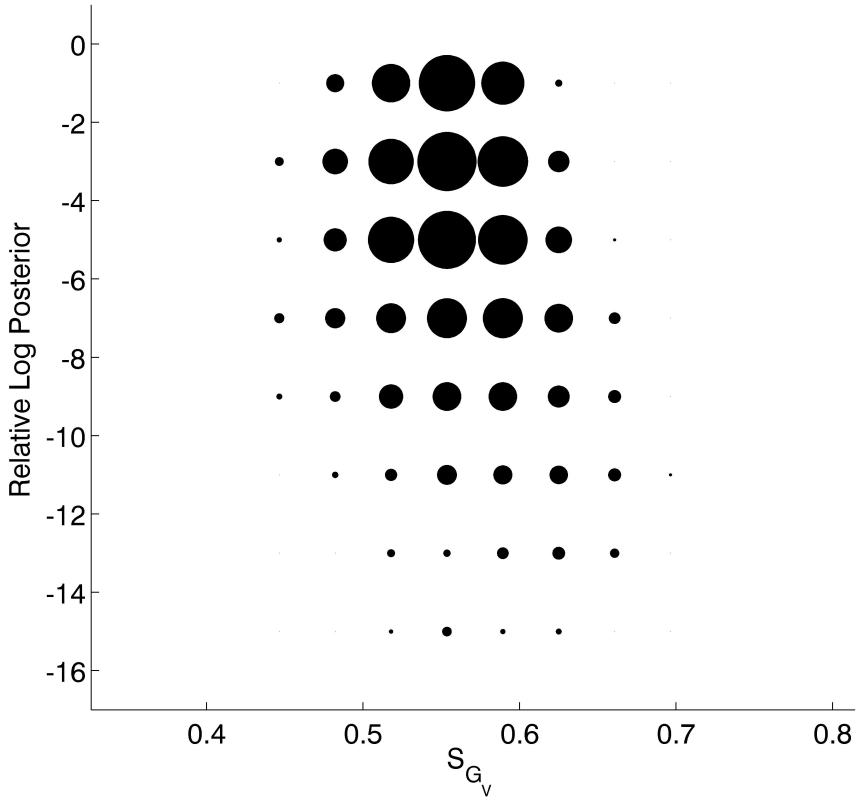

FIGURE 3.5: Summary of posterior on sparsity of $G_{\mathbf{V}}$ in the econometric example. Circle areas are proportional to the fraction of posterior sampled graphs at several levels of posterior probability plotted against levels of sparsity $S_{G_{\mathbf{V}}}$ measured as the proportion of edges included.

itly constrained hyper-inverse Wishart prior is easy to interpret, specify and, with results from Carvalho et al. (2007), implement; synthetic examples have verified the resulting efficacy of the simulation and model search computations.

Our use of Candidate's formula to provide different approximations to marginal likelihoods over graphs can be extended to multiple such approximations. I have explored other constructions, and find no obvious practical differences in resulting estimates in simulated examples. This is an area open for theoretical investigation and in other model contexts. This also offers a route to extending the analysis here to non-decomposable graphical models.

Examples in Section 3.4.1 and 3.4.2 are in modest dimensional problems where local move Metropolis-Hastings methods for the graphical model components of the analysis can be expected to be effective, building on experiences in multivariate models (Jones et al., 2005). To scale to higher dimensions, alternatives computational strategies such as shotgun stochastic search over graphs (Dobra et al., 2004; Jones et al., 2005; Hans et al., 2007) become relevant. A critical perspective is to define analysis that will rapidly find regions of graphical model space supported by the data. It is far better to work with a small selection of high-probability models than a grossly incorrect model on full graphs, and as dimensions scale the latter quickly becomes infeasible. Shotgun stochastic search and related methods reflect this and offer a path towards faster, parallelisable model search. There is also potential for computationally faster approximations using expectation-maximisation style and variational methods (Jordan et al., 1999). The example in Section 3.2 is illustrative of the statistical and computational methodology in a higher-dimensional problem, while introducing a novel class of Markov random field models that emerge quite naturally from the matrix graphical model context. With the matrix data representing a spatial process on a rectangular grid, taking covariance matrices $\mathbf{U}$ and $\mathbf{V}$ as those of two stationary autoregressive processes provides flexibility in modelling patterns separately in horizontal and vertical directions.

4

Dynamic financial index models: Modeling conditional dependencies via graphs

4.1 Introduction

Since the seminal work of Sharpe (1964), Financial Index Models have been in the core of asset pricing and portfolio allocation problems. These models assume that all systematic variation in the returns of financial securities can be explained by one or a set of market indices (factors). The central empirical implication of this assumption is a highly structured covariance matrix for the distribution of returns as, after conditioning on the chosen set of market indices, the residual covariance matrix is diagonal. The attractiveness of this approach is immediate as it offers a very simple, economically justifiable and stable way to estimate potentially very large covariance matrices.

The covariance matrix of returns is a key input in building optimal portfolios and its estimation is often challenging as the number of parameters grows exponentially with the number of assets considered. It is necessary, therefore, to work with structured models that reduce the dimensionality of the problem and deliver

more effective estimates and, in turn, better investment decisions. In this chapter, I explore a generalisation of Financial Index Models with more complex patterns of covariation between returns by allowing conditional dependencies via the introduction of graphical constraints. I work with the matrix-variate dynamic graphical model (DGM) framework of Carvalho & West (2007a,b) but, unlike their original work, graphs are used to increase complexity and not to reduce it.

I take the view that, given its popularity in empirical finance, Index Models such as the *Capital Asset Pricing Model* (CAPM) and the *Fama-French* (FF) are appropriate for the purpose of asset allocation. The central idea of my work it to show that it is possible to improve upon traditional estimates from Index Models and provide more flexible, efficient and still parsimonious strategies for estimating covariances. In addition, I provide two extensions to DGMs: *(i)* I consider the problem of sequential inference about the graphical structure and, *(ii)* define the sequential updating process in the presence of stochastic regressors.

The proposed forecasting model is tested on stock returns data in a portfolio selection exercise. Using 100 NYSE monthly stock returns from 1989 through 2008, I find that our strategy yields better out-of-sample forecast of realised covariance matrix and lower portfolio variance than the two traditional implementations of index models, the capital asset pricing model (CAPM) and the Fama-French (FF) model.

I start by describing Index Models in Section 4.2 along with their use in the dynamic linear model context. Section 4.3 presents the necessary background of dynamic matrix-variate graphical models. In Section 4.4, I discuss issues of dealing with graph (model) uncertainty through time and a simulation study in presented in Section 4.5. Section 4.6 expands the DGM context to allow for random regressors. Finally, in Section 4.7 I explore the use of DGMs as a tool to improve the implementation of Financial Index Models.

4.2 Financial Index Models

A k-dim Index Model assumes that stock returns are generated by

$$Y_{it} = \alpha_i + \sum_{j=1}^{k} \theta_{ij} f_{jt} + \nu_{it}$$

where f_{jt} is the jth common factor at time t, and residuals ν_{it} are uncorrelated to index f_{jt} and to one another. This implies that the covariance matrix of returns can be written as:

$$\mathbf{V}_t = \mathbf{\Theta}' \mathbf{\Psi}_t \mathbf{\Theta} + \mathbf{\Sigma}_t$$

where $\mathbf{\Theta}$ is the matrix of factor loadings of stocks, $\mathbf{\Psi}_t$ is the covariance matrix of the factors, and $\mathbf{\Sigma}_t$ is a diagonal matrix containing the residual return variances.

Some interesting Index Models include the single index model and three index model. The single index uses the excess return of the market as the single index. This model corresponds to the standard Capital Asset Pricing Model of Sharpe (1964). More recently, and perhaps the most commonly used approach is the three index model proposed by Fama & French (1993) where two new factors (besides the market) are added: value-weighted market index with size and book-to-market factors.

These models are usually estimated by running a set of independent regressions where the excess return of each stock is regressed against the indices for a certain window of time. Call $\hat{\boldsymbol{\theta}}_i$ the estimates of the regression coefficients for stock i and $\hat{\sigma}_{ii}$ the residual variance estimate. This yields the following estimator for the covariance matrix of stock returns:

$$\hat{\mathbf{V}} = \hat{\mathbf{\Theta}} \hat{\mathbf{\Psi}} \hat{\mathbf{\Theta}}' + \hat{\mathbf{\Sigma}},$$

where $\hat{\mathbf{\Psi}}$ is the sample covariance matrix of indices, $\hat{\mathbf{\Theta}} = [\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_p]$ is the matrix of regression coefficients for all p assets and $\hat{\mathbf{\Sigma}}$ is the diagonal matrix of residual

variances. This strategy usually defines the one-step forecast of the covariance matrix to be the current estimate of the covariance matrix $\hat{\mathbf{V}}$.

In our work, I recast the above strategy in a natural model based on a state-space or dynamic linear model (DLM) (West & Harrison, 1997) representation. This follows the work of Zellner & Chetty (1965); Quintana & West (1987); Carvalho & West (2007a), to cite a few. I use a dynamic regression framework where, in its full generality, a $p \times 1$ vector time series of returns $\mathbf{Y}_t$ follows the dynamic linear model

$$\mathbf{Y}_t' = \mathbf{F}_t'\boldsymbol{\Theta}_t + \boldsymbol{\nu}_t', \qquad \boldsymbol{\nu}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{t}), \tag{4.1}$$

$$\boldsymbol{\Theta}_t = \boldsymbol{\Theta}_{t-1} + \boldsymbol{\Omega}_t, \qquad \boldsymbol{\Omega}_t \sim N(\mathbf{0}, \mathbf{W}_t, \boldsymbol{\Sigma}_\mathbf{t}), \tag{4.2}$$

for $t = 1, 2, \ldots,$ where

(a) $\mathbf{Y}_t = (Y_{ti})$, the $p \times 1$ observation vector;

(b) $\boldsymbol{\Theta}_t = (\boldsymbol{\theta}_{ti})$, the $n \times p$ matrix of states;

(c) $\boldsymbol{\Omega}_t = (\boldsymbol{\omega}_{ti})$, the $n \times p$ matrix of evolution innovations;

(d) $\boldsymbol{\nu}_t = (\nu_{ti})$, the $p \times 1$ vector of observational innovations;

(e) for all t, the $n \times 1$ regressor vector $\mathbf{F}_t$, is known.

Also, $\boldsymbol{\Omega}_t$ follows a matrix-variate normal with mean $\mathbf{0}$, left and right covariance matrices $\mathbf{W}_t$ and $\boldsymbol{\Sigma}_\mathbf{t}$, respectively. In terms of scalar elements, I have p univariate models with individual n-vector state parameters, namely

$$\text{Observation:} \quad Y_{ti} = \mathbf{F}_t'\boldsymbol{\theta}_{ti} + \nu_{ti}, \qquad \nu_{ti} \sim N(0, \sigma^2_{ii,t}), \tag{4.3}$$

$$\text{Evolution:} \quad \boldsymbol{\theta}_{ti} = \boldsymbol{\theta}_{t-1,i} + \boldsymbol{\omega}_{ti}, \qquad \boldsymbol{\omega}_{ti} \sim N(\mathbf{0}, \mathbf{W}_t\sigma^2_{ii,t}), \tag{4.4}$$

for each i, t. Each of the scalar series shares the same $\mathbf{F}_t$ elements, and the reference to the model as one of exchangeable time series reflects these symmetries. This is a

standard specification in which the correlation structures induced by $\mathbf{\Sigma}_t$ affect both the observation and evolution errors; for example, if $\sigma_{ij,t}$ is large and positive, vector series i and j will show concordant behavior in movement of their state vectors and in observational variation about their levels. Specification of the entire sequence of W_t in terms of discount factors (West & Harrison, 1997) is also standard practice, typically using discount factors related to the state vector and their expected degrees of random change in time.

The above representation provides sequential, closed-form analytical updates of the one-step ahead forecast distributions of future returns and posterior distributions for states and parameters defining the model. This allows for proper accounting of the uncertainty associated with all necessary inputs in sequential investment decisions.

According to traditional Index Models, $\mathbf{\Sigma}_t$ is a diagonal matrix as all common variation between returns should be captured by the elements in $\mathbf{\Theta}_t$. I will depart from this standard assumption and allow for a more flexible representation of the residual covariance matrix leading to potentially more complex forms of $\mathbf{V}$. This is done via the introduction of conditional independencies determined by graphical constraints in $\mathbf{\Sigma}_t$. The use of these models in sequential portfolio problems is originally proposed by Carvalho & West (2007a) and further analyzed by Quintana et al. (2009). In both references however, graphs were used to reduce the dimensionality of an otherwise fully unstructured covariance matrix of returns. Here, I come from a different direction and show that graphs can be successfully used to increase the complexity of an otherwise highly structured covariance matrix. Before continuing, I need to define the necessary notation for the introduction of graphical models in DLMs.

4.3 Dynamic matrix-variate graphical model

The matrix-variate graphical model framework combines hyper-inverse Wishart prior distributions of equation (1.3) together with matrix and multivariate normal distributions, in a direct and simple extension of the usual normal, inverse Wishart distribution theory to the general framework of graphical models. The $n \times p$ random matrix $\mathbf{X}$ and $p \times p$ random variance matrix $\mathbf{\Sigma}$ have a joint matrix-normal, hyper-inverse Wishart (NHIW) distribution if $\mathbf{\Sigma} \sim HIW_G(b, \mathbf{D})$ on G and $(\mathbf{X}|\mathbf{\Sigma}) \sim N(\mathbf{m}, \mathbf{W}, \mathbf{\Sigma})$ for some $b, \mathbf{D}, \mathbf{m}, \mathbf{W}$. I denote this by $(\mathbf{X}, \mathbf{\Sigma}) \sim NHIW_G(\mathbf{m}, \mathbf{W}, b, \mathbf{D})$ with $\mathbf{X}$ marginally following a matrix hyper-T (as defined in Dawid & Lauritzen, 1993) denoted by $HT_G(\mathbf{m}, \mathbf{W}, \mathbf{D}, b)$.

In the dynamic linear model context and given $\mathbf{\Sigma}_t$ constrained by any decomposable graph G, Carvalho & West (2007a,b) define the details of the full sequential and conjugate updating, filtering and forecasting for the dynamic regressions and time-varying $\mathbf{\Sigma}_t$. This approach incorporates graphical structuring into the traditional matrix-variate DLM context and provides a parsimonious yet tractable model for $\mathbf{\Sigma}_t$. Consider the matrix normal DLM described in equation (4.1) and (4.2). With the usual notation that $D_t = \{D_{t-1}, \mathbf{Y}_t\}$ is the data and information set upon any time t, assume the NHIW initial prior of the form

$$(\mathbf{\Theta}_0, \mathbf{\Sigma}_0 \mid D_0) \sim NHIW_G(\mathbf{m}_0, \mathbf{C}_0, b_0, \mathbf{S}_0). \tag{4.5}$$

In components, $(\mathbf{\Theta}_0 \mid \mathbf{\Sigma}_0, D_0) \sim N(\mathbf{m}_0, \mathbf{C}_0, \mathbf{\Sigma}_0)$ and $(\mathbf{\Sigma}_0 \mid D_0) \sim HIW_G(b_0, \mathbf{S}_0)$, which incorporates the conditional independence relationships from G into the prior. For now assume full knowledge of G defining the conditional independence relationships in $\mathbf{Y}$. Full sequential updating can be summarised as follows:

Theorem 1. *(Carvalho & West, 2007a,b) Under the initial prior of equation (4.5) and with data observed sequentially to update information sets D_t the sequential*

updating for the matrix normal dynamic graphical models (DGM) on G is given as follows:

(i) *Posterior at* $t-1$*:* $(\mathbf{\Theta}_{t-1}, \mathbf{\Sigma}_{t-1} \mid D_{t-1}) \sim NHIW_G(\mathbf{m}_{t-1}, \mathbf{C}_{t-1}, b_{t-1}, \mathbf{S}_{t-1})$

(ii) *Prior at* t*:* $(\mathbf{\Theta}_t, \mathbf{\Sigma}_t \mid D_{t-1}) \sim NHIW_G(\mathbf{a}_t, \mathbf{R}_t, \delta b_{t-1}, \delta \mathbf{S}_{t-1})$ *where* $\mathbf{a}_t = \mathbf{m}_{t-1}$ *and* $\mathbf{R}_t = \mathbf{C}_{t-1} + \mathbf{W}_t$

(iii) *One-step forecast:* $(\mathbf{Y}_t \mid D_{t-1}) \sim HT_G(\mathbf{f}_t, q_t \delta \mathbf{S_{t-1}}, \delta b_{t-1})$ *where* $\mathbf{f}'_t = \mathbf{F}'_t \mathbf{a}_t$ *and* $q_t = \mathbf{F}'_t \mathbf{R}_t \mathbf{F}_t + 1$

(iv) *Posterior at* t*:* $(\mathbf{\Theta}_t, \mathbf{\Sigma}_t \mid D_t) \sim NHIW_G(\mathbf{m}_t, \mathbf{C}_t, b_t, \mathbf{S}_t)$ *with* $\mathbf{m}_t = \mathbf{a}_t + \mathbf{A}_t \mathbf{e}'_t$, $\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t \mathbf{A}' q_t$, $b_t = \delta b_{t-1} + 1$, $\mathbf{S}_t = \delta \mathbf{S}_{t-1} + \mathbf{e}_t \mathbf{e}'_t / q_t$ *where* $\mathbf{A}_t = \mathbf{R}_t \mathbf{F}_t / q_t$ *and* $\mathbf{e}_t = \mathbf{Y}_t - \mathbf{f}_t$.

The above derivation uses a "locally smooth" discount factor-based model to allow $\mathbf{\Sigma}_t$ to vary stochastically. This is a common approach in dynamic linear models (Quintana et al., 2003) where information is discounted through time by a pre-specified discount factor δ. This provides sequential estimates of $\mathbf{\Sigma}_t$ that keep adapting to new data while further discounting past observations. This is easily seen in the representation of the posterior harmonic mean that has the form of a exponentially weighted moving average estimate define as

$$\hat{\mathbf{\Sigma}}_t \approx (1-\delta) \sum_{l=0}^{t-1} \delta^l \mathbf{e}_{t-l} \mathbf{e}'_{t-l}.$$

In practical terms the choice of δ represents a similar problem as the choice of the data window in the usual estimation of index models. Extensive discussion of choice of δ in dynamic variance models appears in Chapter 16 of West & Harrison (1997).

So far, G was assumed known and held fixed for all t. This is clearly a limitation of the framework of Carvalho & West (2007a) as it is not necessarily the case

that the same set of conditional independence constraints remain fixed across time. Moreover, it is rarely the case that knowledge about G is available and data driven approaches to determine G are required which represents a non-trivial question in empirical applications. Carvalho & West (2007a) present one example where graphs were selected via the computationally intensive stochastic search ideas of Jones et al. (2005). Quintana et al. (2009) consider similar strategies and briefly explore the issue of time variation in G when modeling currencies. It is clear that the use of dynamic matrix-variate graphical models requires proper accounting of the uncertainty associated with G. Before continuing in our exploration of the use of graphs in index models, I add to this discussion and consider alternatives to learn about the conditional independence relationships defining the models.

4.4 Graphical model uncertainty and search

4.4.1 Marginal likelihood over graphs

In the standard static context, model selection involves the posterior distribution of graphs, given by:

$$p(G \mid \mathbf{x}) \propto p(\mathbf{x} \mid G)p(G)$$

where $p(\mathbf{x} \mid G)$ is the marginal likelihood of G. The marginal likelihood function for any graph G is computed by integrating out the covariance matrix with respect to the prior

$$p(\mathbf{x} \mid G) \;=\; \int_{\mathbf{\Sigma}^{-1} \in M(G)} p(\mathbf{x} \mid \mathbf{\Sigma}, G)p(\mathbf{\Sigma} \mid G)\mathrm{d}\mathbf{\Sigma}$$

where $M(G)$, as before,indicates the set of all positive-definite symmetric matrices constrained by G.

Under a hyper-inverse Wishart prior for $\mathbf{\Sigma}$ and observed data $\mathbf{x}$ of sample size n, the above integration for decomposable graph becomes a simple function of the

prior and posterior normalising constants, $H(b, \mathbf{D}, G)$ and $H(b+n, \mathbf{D}+\mathbf{S}_{\mathbf{x}}, G)$:

$$p(\mathbf{x} \mid G) \;=\; (2\pi)^{-np/2} \frac{H(b, \mathbf{D}, G)}{H(b+n, \mathbf{D}+\mathbf{S}_{\mathbf{x}}, G)}$$

where the normalising constant $H(b, \mathbf{D}, G)$ is given by

$$H(b, \mathbf{D}, G) \;=\; \frac{\prod_{P \in \mathcal{P}} |\frac{\mathbf{D}_P}{2}|^{(\frac{b+|P|-1}{2})} \Gamma_{|P|}(\frac{b+|P|-1}{2})^{-1}}{\prod_{S \in \mathcal{S}} |\frac{\mathbf{D}_S}{2}|^{(\frac{b+|S|-1}{2})} \Gamma_{|S|}(\frac{b+|S|-1}{2})^{-1}}, \tag{4.6}$$

with $\Gamma_k(a)$ the multivariate gamma function.

In the dynamic set-up, a fully Bayesian analysis will consider the graph predictive probability of $\pi(G \mid D_{t-1})$ over $\mathcal{G}$, the set of all decomposable graphs, and specify the unconditional predictive distribution $p(\mathbf{Y}_t \mid D_{t-1})$ as $\mathrm{E}_G\{p(\mathbf{Y}_t \mid D_{t-1}, G)\}$ with the expectation taken with respect to $p(G \mid D_{t-1})$, namely,

$$(\mathbf{Y}_t \mid D_{t-1}) \;\sim\; \sum_{G \in \mathcal{G}} \pi(G \mid D_{t-1}) p(\mathbf{Y}_t \mid D_{t-1}, G). \tag{4.7}$$

Equation (4.7) indicates that the predictive probability $\pi(G \mid D_{t-1})$ is central to evaluating the predictive distribution $p(\mathbf{Y}_t \mid D_{t-1})$. The two possibilities for consideration of predicting G are as follows:

(i) fixed graph for all t, that is for some $G \in \mathcal{G}$, DLM(G) holds for all t;

(ii) time varying graphs where for some possible sequence of graphs $G_t \in \mathcal{G}$,($t = 1, 2, \cdots$), DLM(G_t) holds at time t.

For (i), the predictive probability of graphs for time t is defined as

$$\pi(G \mid D_{t-1}) \;=\; p(G \mid D_{t-1}) \propto p(G) p(\mathbf{Y}_{1:t-1} \mid G) \tag{4.8}$$

where the marginal likelihood of a DLM on any graph G is

$$p(\mathbf{Y}_{1:t-1}|G) = p(\mathbf{Y}_{t-1}|D_{t-2}, G) p(\mathbf{Y}_{t-2}|D_{t-3}, G) \ldots p(\mathbf{Y}_1|D_0, G),$$

with each element in the product, $(\mathbf{Y}_t|D_{t-1}, G) \sim HT_G(\mathbf{f}_t, \mathbf{S_{t-1}}, b_{t-1})$ as defined in Theorem 1.

For (ii), the time dependence is made explicit with time subscripts, so that a graph G_i at time t is $G_{t,i}$. Denote $\pi(G_{t,i} \mid D_{t-1})$ as the predictive probability at time $t-1$ for graph G_i. It is natural to dynamic modeling that, as time progresses, what occurred in the past becomes less and less relevant to inference made for the future. Applying this notion to graphs, past data loses relevance to current graphs as t increases. Once again, one practical possibility is to use a discount factor to reduce the impact of past information to current inferences, similarly to the discounting ideas used in modeling $\mathbf{\Sigma}_t$. I propose the following predicted probability of $G_{t,i}$ for time t at time $t-1$

$$\pi(G_{t,i} \mid D_{t-1}) \propto \frac{H(b_0, \mathbf{S}_0, G_{t,i})}{H(\delta b_{t-1}, q_t \delta \mathbf{S}_{t-1}, G_{t,i})} \pi_0(G_{t,i}), \tag{4.9}$$

where δ is the same discount factor as in Theorem 1.

To provide insights into the nature of the predicted probability (4.9), suppose the graph has the prior (4.9) at time $t-1$. Proceeding to observe $\mathbf{Y}_t$, this prior updates to posterior via the usual updating equations:

$$\begin{aligned} \pi(G_{t,i} \mid D_t) &\propto p(\mathbf{Y}_t \mid D_{t-1}, G_{t,i}) \pi(G_{t,i} \mid D_{t-1}) \\ &\propto \frac{H(\delta b_{t-1}, q_t \delta \mathbf{S}_{t-1}, G_{t,i})}{H(b_t, q_{t+1} \mathbf{S}_t, G_{t,i})} \frac{H(b_0, \mathbf{S}_0, G_{t,i})}{H(\delta b_{t-1}, q_t \delta \mathbf{S}_{t-1}, G_{t,i})} \pi_0(G_{t,i}) \\ &= \frac{H(b_0, \mathbf{S}_0, G_{t,i})}{H(b_t, q_{t+1} \mathbf{S}_t, G_{t,i})} \pi_0(G_{t,i}), \end{aligned} \tag{4.10}$$

which has the same representation as equation (4.9), i.e. a ratio of two normalising constants of hyper-inverse Wishart distributions, but updated location parameter and degrees of freedom, i.e. $\mathbf{S}_t = \delta \mathbf{S}_{t-1} + \mathbf{Y}_t \mathbf{Y}_t'$ and $b_t = \delta b_{t-1} + 1$. On the other hand, by substituting $t+1$ for t in prior (4.9), I obtain the prior probability for

$\pi(G_{t+1,i} \mid D_t)$ at t as follows

$$\pi(G_{t+1,i} \mid D_t) \propto \frac{H(b_0, \mathbf{S}_0, G_{t+1,i})}{H(\delta b_t, q_{t+1}\delta \mathbf{S}_t, G_{t+1,i})}\pi_0(G_{t+1,i}), \tag{4.11}$$

which, in comparison with equation (4.10), has introduced a discount factor δ to model a decay of information between time t and $t+1$ in a way analogues to the standard use of discount factors in DLMs. The maintenance of the normalising constant ratio prior and posterior probability at each time enables continued, easy sequential updating, with the minor modification that the degrees of freedom b_t is discounted successively.

This predicted model of equation (4.9) also implies that, the most recent exponentially weighted residual covariance matrix $\mathbf{S}_{t-1}$ could predict both the one-step ahead residual graphical structure and the residual covariance matrix.

Given any particular graph predicting model M_G and discount factor δ, predicting $\mathbf{Y}_t$ in $\alpha = (M_G, \delta)$ is based on the predictive density

$$p(\mathbf{Y}_t \mid D_{t-1}, \alpha) \;=\; \sum_{G_{t,i} \in \mathcal{G}} p(\mathbf{Y}_t \mid D_{t-1}, \alpha, G_{t,i}) p(G_{t,i} \mid D_{t-1}, \alpha). \tag{4.12}$$

4.4.2 Sequential stochastic search

Regardless of the choice of model for G, the model selection problem is further complicated by the explosive combinatorial nature of the space of possible graphs. Without the restriction of decomposability there are $2^{\binom{p}{2}}$ elements in graph space, where p represents the number of vertices. Decomposability accounts for approximately 10% of this number which is still impossible to enumerate for moderate size p. Any attempt to deal with these models requires the development of efficient computational tools to explore the model space. Here, I propose an extension to the shotgun stochastic search (SSS) of Jones et al. (2005) to sequentially learn $(G_{t,i} \mid D_{t-1})$. In a nutshell,

our analysis generates multiple graphs at each time t from the predictive probability $\pi(G_{t,i} \mid D_{t-1})$, using SSS.

Suppose that, at time $t-1$, I have saved a sample of the top N graphs $G_{t-1,i}, i = 1, \cdots, N$ with highest predictive probabilities $\pi(G_{t,i} \mid D_{t-1})$. Proceeding to time t, I adopt the following search algorithm:

(i) Evaluate new predictive probabilities $\pi(G_{t+1,i} \mid D_t)$ of these N graphs from time $t-1$;

(ii) From among the N graphs, propose the ith graph as a new starting graph with probability proportional to $\pi(G_{t+1,i} \mid D_t)^c$, where c is an annealing parameter;

(iii) Start with $G_{t+1,i}$ and apply SSS. After each stage of SSS, compute the Bayesian model average (BMA) estimator of a predicted quantity of interest, e.g. predictive covariance matrix, using current top N graphs;

(iv) Stop search when certain distance between the last two BMA estimates is below a small number, set $t = t+1$ and return to (i).

The evaluation and resample steps of (i) and (ii) are important because "top graphs" from the previous step still represent the majority of our knowledge and should be good starting points for a new SSS once a new data sample becomes available.

4.5 An example

To focus the idea of sequential learning in dynamic graphical model, I first consider a local trend DLM, namely

$$\begin{aligned} \mathbf{Y}_t &= \boldsymbol{\theta}_t + \boldsymbol{\nu}_t, \qquad \boldsymbol{\nu}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_t), \\ \boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(\mathbf{0}, W_t\boldsymbol{\Sigma}_t). \end{aligned}$$

This is a special case of the general DLMs presented in previous sections. I extend the example in Carvalho & West (2007a) where data from $p = 11$ international currency exchange rates relative to the US dollar is analyzed. Figure 4.1 shows the time series plots of these 11 exchange rate returns. In all models, I use fairly diffuse priors, and annealing parameter $c = 1$.

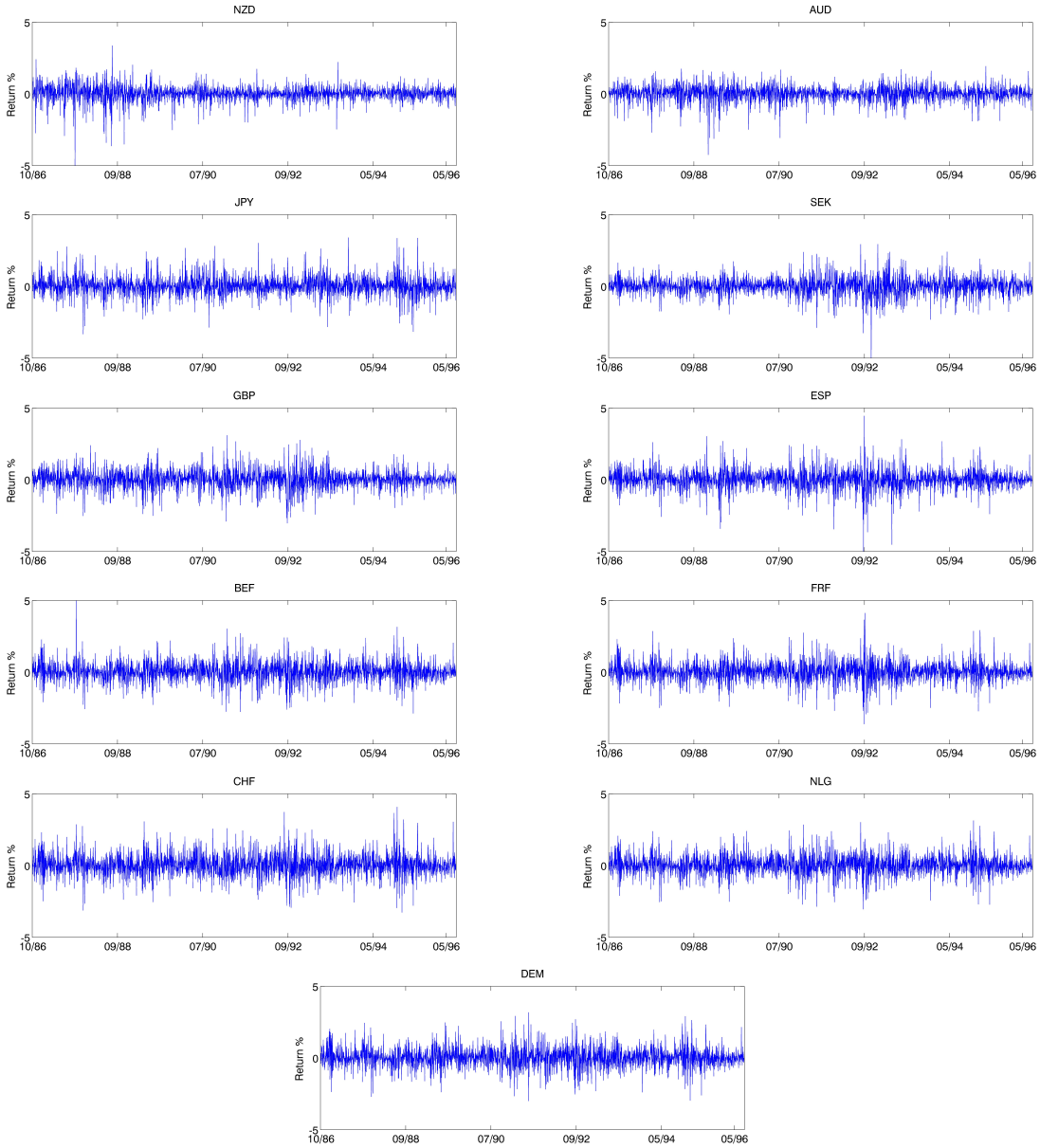

FIGURE 4.1: Time series plots of daily exchange rate returns.

For illustration, I ran a set of parallel analyses for eight different combinations of the discount factor and the graph predicting model. In particular, let M_G be the graph predicting model that takes its value in the set $\{M_F, M_C\}$ where M_F represents the fixed graph predicting model as is described by equation (4.8) and M_C represents the time-varying graph predicting model as is described by equation (4.9). Let δ take its value in the set $\{0.93, 0.95, 0.97, 0.99\}$. Then at each of the eight pairs of (M_G, δ), and time t, the marginal likelihood of equation (4.12) is approximated by summing over top 1000 graphs at each time t, resulting in a full marginal likelihood function of (M_G, δ).

Figure 4.2 displays the plots over time of log Bayes factors for each of the eight models against the model $(M_F, 0.95)$. When comparing Bayes factors within each δs, Figure 4.2 shows that all four time-varying graphs generate smaller marginal likelihood values as their fixed graph peers. Figure 4.3 highlights the change of the log Bayes factors of the top two models. Overall, the chosen MLE from such analysis is $(M_F, 0.97)$ over the period up to the end of 08/1992 and $(M_F, 0.95)$ over the period from then until the end of data at 06/1996. The change from $\delta = 0.97$ to $\delta = 0.95$ at the end of 08/1992 reflects a more adaptive model favoured later. The occurrence of one or two rather marked changes may be due to major economic changes and events. A key such event was Britain's withdrawal from the EU exchange rate agreement (ERM) in the September 1992 and into 1993 that led to the deviation from the steady behaviour anticipated under a model with relative high discount factor 0.97 to the more adaptive 0.95. A second period of change of structure occurred in early 1995 with major changes in Japanese interest rate policies as a response to a weakening Yen and a move toward financial restructuring in Japan.

Figure 4.4 displays four snapshots of predicted adjacency matrices at four different time points. These adjacency matrices are from the median probability graph (Scott & Carvalho, 2008), which is defined as the graph consisting of those edges whose

overall edge inclusion probability exceeds 0.5. At each time t, the edge inclusion probabilities are predicted from the outputs of the graphical model search under top identified models: $(M_F, 0.97)$ before 09/1992 and $(M_F, 0.95)$ after 09/1992. As can be seen, these best predicted graphs have several persistent signals as well as the similar overall patterns over time.

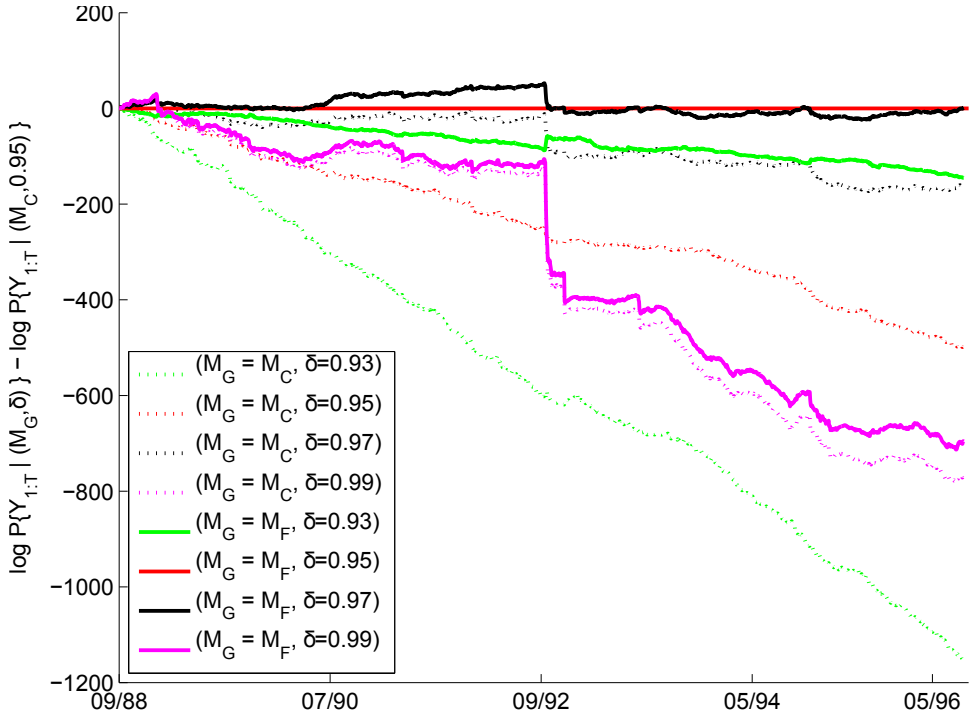

FIGURE 4.2: Log Bayes factors for each of the eight models against the model $(M_F, 0.95)$. The eight models represent the eight different combinations of δ from four distinct values $\{0.93, 0.95, 0.97, 0.99\}$ and M_G from two distinct graph predicting models: M_F, the fixed graph predicting model as is described by equation (4.8); M_C, the time-varying graph predicting model as is described by equation (4.9). The figure illustrates that the marginal likelihoods of time-varying graphs (dashed lines) are generally smaller than those of fixed graphs (solid lines)

This example serves to illustrate some features of inference with dynamic graphical models. In each of the DGMs (M_G, δ), and for any specified sequence of graphs $\{G_t\}$, the prior, posterior, and forecast distributions are all standard distributions of well-understood forms, whether they be hyper-inverse Wishart or hyper T. Fore-

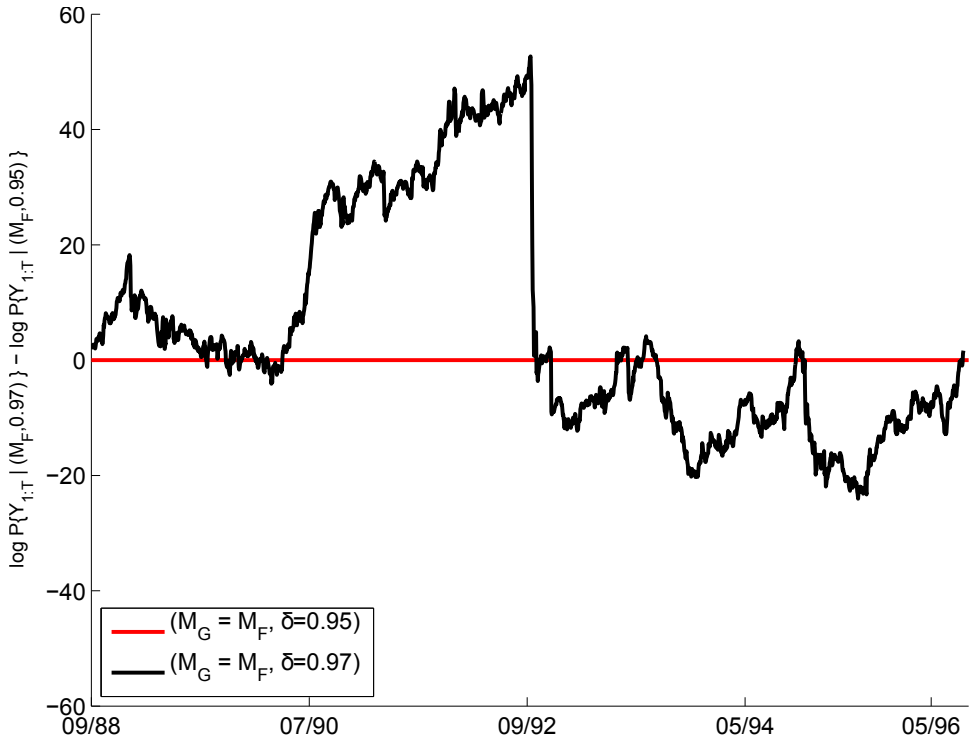

FIGURE 4.3: Log Bayes factors for the model $(M_F, 0.97)$ against the model $(M_F, 0.95)$

casts that take into account uncertainty about graphs are easily calculated from the finite mixture of hyper T distributions of equation (4.12). If one is concerned about which are the best graph predicting models or which discount factors to use, their corresponding Bayes factors may be used to choose these specifications.

The two proposed graph predicting models together with the covariance matrix discount factors allows us to separately infer the dynamics of graphs and the dynamics of covariance matrices. In this particular example the marginal likelihoods favor static models M_F for all values of δ's. This suggests that time-varying graphs inferred by a moving window may not produce consistently better predicting results than fixed graphs with signals detected sequentially using all historical data. On the other hand, covariance matrices $\mathbf{\Sigma}_t$ seem to be time-varying, since the marginal likelihoods of δ favor values of 0.95 or 0.97. Further exploration of these issues

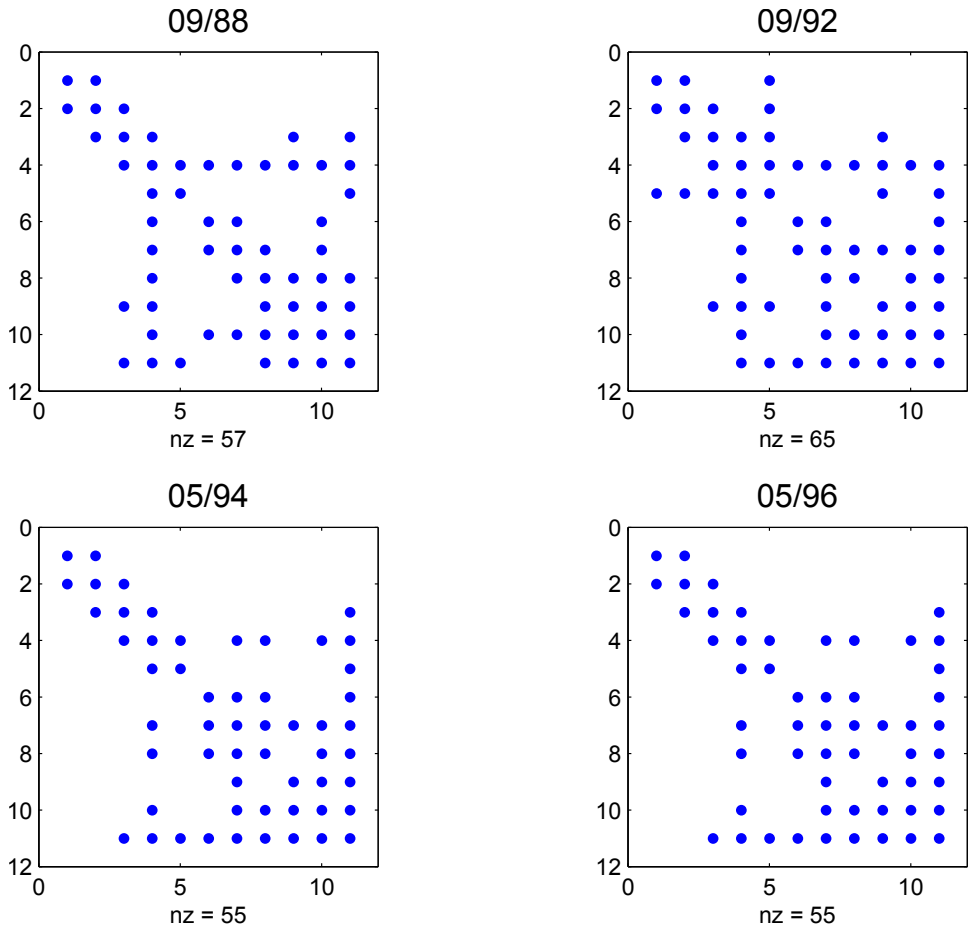

FIGURE 4.4: Four snapshots of adjacency matrices of median probability graphs predicted by using the output the stochastic search under corresponding top models: $(M_F, 0.97)$ (upper two panels), and model $(M_F, 0.95)$ (lower two panels).

are necessary and should be regarded as applied questions to be dealt with in any a particular application. My aim here has been to demonstrate that the methods presented in Section 4.4 provide a computationally attractive way to address these modeling questions.

4.6 Random regression DLM

Applied interests are motivated by models where I attempt to predict $\mathbf{Y}_t$ with a regression vector $\mathbf{F}_t$ that is random and unknown before time t. Now, let $I_t =$

$\{\mathbf{Y}_1, \cdots, \mathbf{Y}_t, \mathbf{F}_t, \cdots, \mathbf{F}_t\}$ denote the data and information set. Assume $\mathbf{F}_t$ has a prior $p(\mathbf{F}_t \mid I_{t-1})$ at time t. Then under the assumption that the priors of $(\mathbf{\Theta}_t, \mathbf{\Sigma}_t)$ and $\mathbf{F}_t$ are conditionally independent given I_{t-1}, namely, $(\mathbf{\Theta}_t, \mathbf{\Sigma}_t) \perp\!\!\!\perp \mathbf{F}_t \mid I_{t-1}$, the following results apply.

Theorem 2. *Under the initial prior of equation (4.5) and with data observed sequentially to update information sets I_t the sequential updating for the matrix normal DLM on G is given as follows:*

(i) *Posterior at $t-1$:* $(\mathbf{\Theta}_{t-1}, \mathbf{\Sigma}_{t-1} \mid I_{t-1}) \sim NHIW_G(\mathbf{m}_{t-1}, \mathbf{C}_{t-1}, b_{t-1}, \mathbf{S}_{t-1})$;

(ii) *Prior at t:* $(\mathbf{\Theta}_t, \mathbf{\Sigma}_t \mid I_{t-1}) \sim NHIW_G(\mathbf{a}_t, \mathbf{R}_t, \delta b_{t-1}, \delta \mathbf{S}_{t-1})$ *where* $\mathbf{a}_t = \mathbf{m}_{t-1}$ *and* $\mathbf{R}_t = \mathbf{C}_{t-1} + \mathbf{W}_t$;

(iii) *One-step forecast:* $p(\mathbf{Y}_t \mid I_{t-1}) = \int HT_G(\mathbf{f}_t, q_t \delta \mathbf{S_{t-1}}, \delta b_{t-1}) p(\mathbf{F}_t \mid I_{t-1}) \mathrm{d}\mathbf{F}_t$ *with first two moments:*

$$\begin{aligned} \mathbf{r}_t &\equiv E(\mathbf{Y}_t \mid I_{t-1}) = \mathbf{a}_t' \mu_{\mathbf{F}_t}, \\ \mathbf{Q}_t &\equiv \mathrm{cov}(\mathbf{Y}_t \mid I_{t-1}) = \mathbf{a}_t' \mathbf{\Sigma}_{\mathbf{F}_t} \mathbf{a}_t + \{V_t + \mu_{\mathbf{F}_t}' \mathbf{R}_t \mu_{\mathbf{F}_t} + \mathrm{tr}(\mathbf{R}_t \mathbf{\Sigma}_{\mathbf{F}_t})\} E(\mathbf{\Sigma}_t \mid I_{t-1}), \end{aligned}$$

where $\mathbf{f}_t' = \mathbf{F}_t' \mathbf{a}_t$ *and* $q_t = \mathbf{F}_t' \mathbf{R}_t \mathbf{F}_t + V_t$, *the first and second moments of the predictive regression vector,* $\mu_{\mathbf{F}_t} = E(\mathbf{F}_t \mid I_{t-1})$ *and* $\mathbf{\Sigma}_{\mathbf{F}_t} = \mathrm{cov}(\mathbf{F}_t \mid I_{t-1})$.

(iv) *Posterior at t:* $(\mathbf{\Theta}_t, \mathbf{\Sigma}_t \mid I_t) \sim NHIW_G(\mathbf{m}_t, \mathbf{C}_t, b_t, \mathbf{S}_t)$ *with* $\mathbf{m}_t = \mathbf{a}_t + \mathbf{A}_t \mathbf{e}_t'$, $\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t \mathbf{A}' q_t$, $b_t = \delta b_{t-1} + 1$, $\mathbf{S}_t = \delta \mathbf{S}_{t-1} + \mathbf{e}_t \mathbf{e}_t'/q_t$ *where* $\mathbf{A}_t = \mathbf{R}_t \mathbf{F}_t / q_t$ *and* $\mathbf{e}_t = \mathbf{Y}_t - \mathbf{f}_t$.

Proof. (i)(ii)(iv) follow directly from Theorem 1. I show the proof of (iii). Using mixture of $p(\mathbf{Y}_t \mid I_{t-1}, \mathbf{F}_t)$ implies that

$$p(\mathbf{Y}_t \mid I_{t-1}) = \int p(\mathbf{Y}_t \mid I_{t-1}, \mathbf{F}_t) p(\mathbf{F}_t \mid I_{t-1}) \mathrm{d}\mathbf{F}_t,$$

where the first term in the integrand, by assumption, is computed as

$$\begin{aligned} p(\mathbf{Y}_t \mid I_{t-1}, \mathbf{F}_t) &= \int p(\mathbf{Y}_t \mid I_{t-1}, \mathbf{F}_t, \mathbf{\Theta}_t, \mathbf{\Sigma}_t) p(\mathbf{\Theta}_t, \mathbf{\Sigma}_t \mid I_{t-1}, \mathbf{F}_t) \mathrm{d}\mathbf{\Theta}_t \mathrm{d}\mathbf{\Sigma}_t \\ &= \int p(\mathbf{Y}_t \mid I_{t-1}, \mathbf{F}_t, \mathbf{\Theta}_t, \mathbf{\Sigma}_t) p(\mathbf{\Theta}_t, \mathbf{\Sigma}_t \mid I_{t-1}) \mathrm{d}\mathbf{\Theta}_t \mathrm{d}\mathbf{\Sigma}_t \\ &= HT_G(\mathbf{f}_t, q_t \delta \mathbf{S_{t-1}}, \delta b_{t-1}). \end{aligned}$$

The unconditional predictive first moment is calculated as:

$$E(\mathbf{Y}_t \mid I_{t-1}) = E_{\mathbf{F}_t|I_{t-1}}\{E(\mathbf{Y}_t \mid I_{t-1}, \mathbf{F}_t)\} = E_{\mathbf{F}_t|I_{t-1}}(\mathbf{a}_t' \mathbf{F}_t \mid I_{t-1}) = \mathbf{a}_t' \mu_{\mathbf{F}_t},$$

and the predictive covariance matrix is derived as:

$$\begin{aligned} \mathrm{cov}(\mathbf{Y}_t \mid I_{t-1}) &= \mathrm{cov}_{\mathbf{F}_t|I_{t-1}}\{E(\mathbf{Y}_t \mid I_{t-1}, \mathbf{F}_t)\} + E_{\mathbf{F}_t|I_{t-1}}\{\mathrm{cov}(\mathbf{Y}_t \mid I_{t-1}, \mathbf{F}_t)\} \\ &= \mathbf{a}_t' \mathrm{cov}(\mathbf{F}_t \mid I_{t-1}) \mathbf{a}_t + \{V_t + E(\mathbf{F}_t' \mathbf{R}_t \mathbf{F}_t \mid I_{t-1})\} E(\mathbf{\Sigma}_t \mid I_{t-1}) \\ &= \mathbf{a}_t' \mathbf{\Sigma}_{\mathbf{F}_t} \mathbf{a}_t + \{V_t + \mu_{\mathbf{F}_t}' \mathbf{R}_t \mu_{\mathbf{F}_t} + \mathrm{tr}(\mathbf{R}_t \mathbf{\Sigma}_{\mathbf{F}_t})\} E(\mathbf{\Sigma}_t \mid I_{t-1}). \end{aligned}$$

□

The above theorem suggests a two-stage model analysis: first, a model is fitted on low dimensional regression vectors $\{\mathbf{F}_t\}$; second, the fitted model provides the necessary quantities for the dynamic graphical DLMs. Some specific contexts of $\{\mathbf{F}_t\}$ include:

- Pre-fixed regression vector in which the $\mathbf{F}_t$ values are specified in advance by design. This is the assumption made by the standard dynamic linear model, which yields a degenerated prior distribution $p(\mathbf{F}_t \mid I_{t-1})$ with $\mu_{\mathbf{F}_t} = \mathbf{F}_t$ and $\mathbf{\Sigma}_{\mathbf{F}_t} = \mathbf{0}$. In such cases, Theorem 1 applies as a special case of Theorem 2.

- Independent and identically distributed regression vector in which the $n-$vector $\mathbf{F}_t$ are commonly assumed to be independent and identically distributed from a multivariate normal distribution with mean vector $\mu_{\mathbf{F}}$ and covariance matrix $\mathbf{\Sigma}_{\mathbf{F}}$.

- Dynamic regression vector in which another dynamic model structure could be imposed on the vector process $\{\mathbf{F}_t\}$. For example, in asset pricing models, if $\mathbf{F}_t$ is the market excessive return, a AR-GARCH type of model could be applied.

4.7 Example: Portfolio allocation in stocks

To demonstrate the use of DGMs in the Index Model context I work with 100 stocks randomly selected from the population of domestic commonly traded stocks in the New York Stock Exchange. By selecting a random sample of 100 I hope to reduce potential selection biases. The sample period is from January 1989 to December 2008 in a total of 240 monthly returns. Monthly returns of a one-month Treasury bill is used as the risk-free rate in the computation of the excess returns. Excess returns from the a market weighted basket of all stocks in the AMEX, NYSE and NASDAQ were used as the *market* returns. This index along with the Fama-French three factor return data were obtained from the data library of Professor Kenneth R. French[1]. Summary statistics for the excess returns series are given in the first row in Table 4.1. The median pairwise correlation is 0.159, indicating that there were potentially large payoffs to portfolio diversification.

In an initial exploration of the data I fitted OLS regressions to the returns using capital asset pricing model (CAPM) and Fama-French (FF) models. The second and third row in Table 4.1 shows summary statistics of cross-sectional residual correlations. The generally lower correlations compared with sample correlation suggest that the indexes capture most of the common variation among the securities under consideration. However, there are remaining signals in the residuals as indicated by the maximum and minimum correlations, and these are in precisely the quantities we are aiming to explore relaxing the independence assumption with the inclusion of graphs.

[1] see, http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Table 4.1: Summary statistics of correlations among sampled stocks. First row, summary of sample correlations; second and third row report summary of residual correlations after fitting CAPM and FF models respectively. For each case, at the end of April of each year from 1994 to 2008, pairwise correlations are calculated based on the monthly excess returns over the prior 60 months. Summary statistics are based on the estimated values pooled over all years.

Model	Mean	Std	Min	25th	Median	75th	Max
Sample	0.158	0.166	−0.530	0.046	0.159	0.269	0.836
CAPM	0.040	0.170	−0.594	−0.075	0.036	0.150	0.825
FF	0.014	0.154	−0.557	−0.092	0.011	0.114	0.816

To appreciate the importance and contribution of the use of graphical models, I consider the following alternatives: (1) sample covariance model; (2) Standard dynamic CAPM; (3) Dynamic CAPM with graphs; (4) Standard dynamic FF; (5) Dynamic FF with graphs, and (6) mixtures of (3) and (5).

In model (1), at each month t, the one-step ahead covariance matrix is based on the data from the preceding 60 months as the *in-sample* period. For model (2)-(6), I use weak priors, $\mathbf{m}_0 = \mathbf{0}, C_0 = 10000\mathbf{I}, b_0 = 3, S_0 = 0.0003\mathbf{I}_{100}$, and $\delta = 0.983$ corresponding to a rolling window of about 60 months.

For the random regression vector $\mathbf{F}_t$, I use sample mean and covariance matrix of the past 60 months as forecasts of the first and second predictive moments, $\boldsymbol{\mu}_{\mathbf{F}_t}$ and $\boldsymbol{\Sigma}_{\mathbf{F}_t}$. Furthermore, based on simulation experiments in Section 4.5, I chose model graph uncertainty with the predictive model of equation (4.8) for alternatives (3) and (5). In (6), CAPM and FF models are compared with each other and then averaged based upon their conditional marginal likelihood $p(\mathbf{Y}_{1:t} \mid \mathbf{F}_{1:t})$. The resulting posterior probabilities of FF model reaches 1 after a short period time. This should not be surprising as most of the current literature points to the use of a multi-factor model as oppose to the traditional single factor CAPM. Due to this fact, the overall performances of model (5) and (6) are close so I only report results from model (5) hereafter.

Figure 4.5 displays the estimated expected number of edges over time starting from January 1994 under model (3) and (5). Three results are worth noting here. First, all graphs are sparse relative to the total 4950 possible edges. The inclusion of graphs provides the necessary flexibility to capture the remaining signals from the residual covariance matrix and the data is responsible to inform which of these non-zero entries are relevant. Second, when comparing with each other, the CAPM model has more edges than FF – once again no surprises here: FF imposes a richer structure for $\mathbf{\Sigma}$ so I should expect more non-zero elements in the residual covariation of assets when the market returns are the only covariate. Third, as more information becomes available, more signals in the residuals are detected.

I now evaluate these forecasting models in two ways: forecasting ability of future correlation matrices and in the construction of optimal portfolios. This is a predictive test, in the sense that our investment strategy does not require any hindsight.

4.7.1 Out-of-sample covariance forecasts

At the end of every month, the correlations forecasts from each model are compared to the sample correlations realised over a subsequent 12 months period, in the first experiment, and 36 months in the second experiment. Forecast performance is evaluated in terms of the absolute difference between the realised and forecasted values. Table 4.2 provide summary statistics on the absolute differences from these two experiments. When the performances evaluated using subsequent 12 months data are compared with those from subsequent 36 months, the average absolute forecast errors are reduced. The drop in forecast errors suggests that there is a lot of noise in covariance matrices measured over a period as short as 12 months. Nevertheless, as in both the 12 and the 36 month experiments, the relative performance of each model are generally the same.

The full sample covariance model, which is the most complex model in terms of

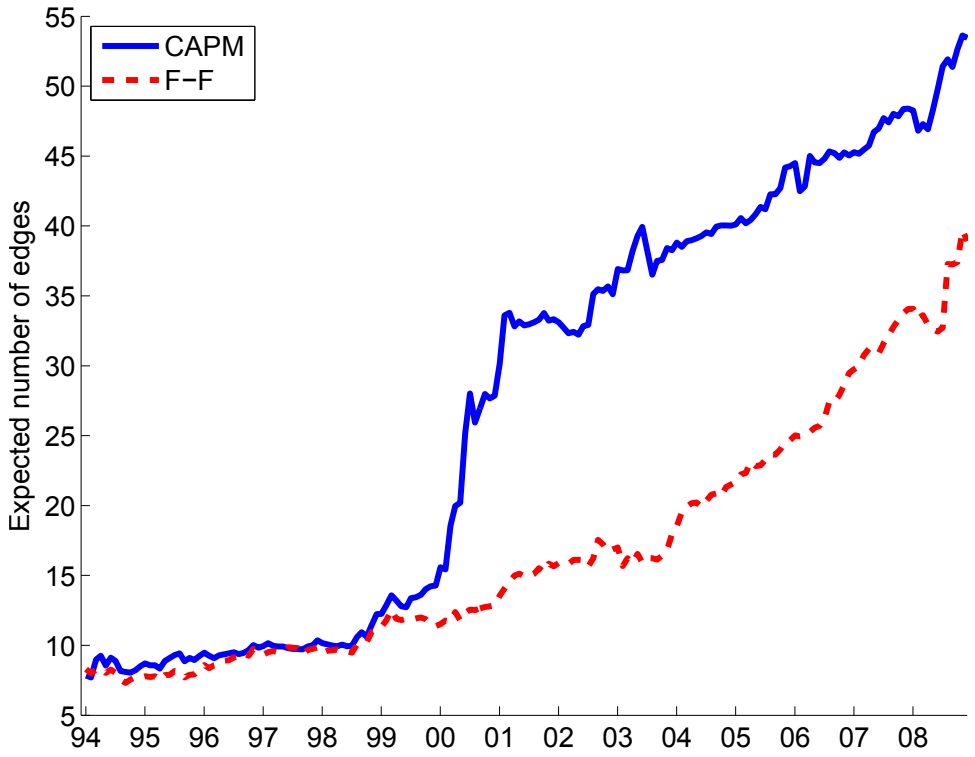

FIGURE 4.5: Estimated expectation of numbers of edge across each month.

number of free parameters, has the highest median absolute error and root mean square error. All other models dominated the full covariance model. More complex models do not necessarily offer smaller forecast errors. This message is consistent with many empirical studies on correlation matrix forecasts of stock returns.

Comparing the empty and the graphical models within either CAPM or FF family, I see that models with graphs dominate their empty graph peers. This is more evident in the CAPM family. Model (3) has reduced the median of the absolute differences and the root mean square errors relative to model (2), while model (5) has almost the same absolute differences as model (4). The clearer advantage in CAPM family is because there is more structure left unexplained in the residuals using only the market index than when using the FF three indexes. In general, the improvement of out-of-sample covariance forecasts is minor. This is actually as expected, since

Table 4.2: Performance of covariance forecasting models. Forecasts of monthly return correlation matrices are generated from different models, based on the prior 60 months of data for model (1) and based on discount factor $\delta = 0.983$ for model (2)-(5). Forecasts are then compared against the realised sample covariance estimated over the subsequent 12 months (first four) columns and 36 months (last four columns). The last estimation period ends in December 2005. Summary statistics are provided for both the distribution of the absolute difference between realised and forecasted value of pairwise correlations: Std, standard deviation of absolute differences; 95th, 95th quantile of absolute difference, and $\sqrt{\text{MSE}}$, the root mean square errors of forecasts.

	12 month				36 month			
Model	Median	Std	95th	$\sqrt{\text{MSE}}$	Median	Std	95th	$\sqrt{\text{MSE}}$
(1) Full covariance	0.238	0.200	0.654	0.340	0.160	0.141	0.460	0.235
(2) CAPM Empty	0.234	0.186	0.612	0.323	0.146	0.127	0.413	0.212
(3) CAPM Graph	0.230	0.184	0.605	0.319	0.143	0.123	0.402	0.207
(4) FF Empty	0.230	0.185	0.609	0.321	0.143	0.124	0.405	0.208
(5) FF Graph	0.230	0.185	0.607	0.320	0.143	0.123	0.404	0.208

the signals are very sparse. However, as the experiments in the following section will show, these signals, though sparse, are influential when the forecast covariance matrices are used to build optimal portfolios.

4.7.2 Portfolio optimisation

From a practical point of view, optimisation experiments provide perhaps more important metrics for evaluating forecasting models. The set-up of our portfolio optimisation experiments is as follows. To highlight the role of the second predictive moment, I first form the global minimum variance portfolio. At the end of April of each year starting from 1994, I use the different models to predict the one-step ahead covariance matrix for the 100 stocks. These predictions are the input to a quadratic programming routine that defines the minimum variance portfolio (Markowitz, 1959). Short sales are allowed so that the weights are only required to be summed up to 1. These weights are then applied to buy-and-hold portfolio returns until the next April,

when the forecasting and optimisation procedures are repeated. The resulting time series of monthly returns of portfolios allow us to characterise the performance of optimised portfolio based on each model. I also form a mean-variance portfolio using the first two moments $\{\mathbf{r}_t, \mathbf{Q}_t\}$ from Theorem 2 with a target annualised excessive mean return of 15%.

Table 4.3 summarises these optimisation results on an annualised basis. In comparison within each group, it is clear that the introduction of the graphical structure helps. The annualised standard deviation of the optimised portfolio based on the graphical CAPM model is 10.7%, yielding a Sharpe ratio of 0.688, compared to a Sharpe ratio of 0.533 for the standard CAPM portfolio. The same advantage of using graphs can be found in two models within FF class. The conclusion from this example is simple: it pays to allow for a more flexible residual covariance structure in the implementation of Index Models.

4.8 Further comments

By allowing more flexible models for the residual covariance matrix, Financial Index Models can be improved in their abilities to build more effective optimal portfolios. In this paper I take advantage of the DGMs framework of Carvalho & West (2007a) and show that graphical models can also be used to identify sparse signals in the residual covariance matrices and thereby obtain a more complex representation of the distribution of asset returns. Unlike Carvalho & West (2007a) and Quintana et al. (2009), in the Index Model framework, graphs are used as a parsimonious way to increase the complexity of an otherwise very restrictive model. In that sense, this our work complements and extends the widely used tool box of dynamic linear models for the analysis asset returns. The first example helps illustrate the model implementation and highlight the issue of specifying discount factors and graph predicting models. The second example discusses and explores aspects of random regression

Table 4.3: Performance of portfolios based on forecasting models. At the end of April of each year from 1994 through 2007, forecasts of covariance matrices of monthly excessive returns are generated from different models. Since $T = 60 < p = 100$, the sample covariance is close to singular, I omit its results. Based upon each model's forecasts of covariance matrices, a quadratic programming procedure is used to find the global minimum variance portfolio (first three column), and mean variance portfolio with target excess annual return 15% (last three column). Short sales are allowed so that the weights are only constrained to sum up to 1. These weights are then applied to form portfolio returns for the next 12 months until next April, at the end of which forecasting and optimisation steps are repeated and the portfolios are formed. Summary statistics are presented: Rate, the annualised excessive returns $r - r_T$, where the annualised portfolio return r is determined by $(1 + r)^{14} = \prod_{i=1}^{168}(1 + r_i)$, and annualised risk-free return r_T is determined by $(1+r_T)^{14} = \prod_{i=1}^{168}(1+r_{T,i})$ with r_i and $r_{T,i}$ denoting the monthly return of portfolio and risk-free asset; Std, the annualised standard deviation of excess returns $r_i - r_{T,i}$; and Sharpe ratio, the annualised excessive return divided by the annualised standard deviation.

	Minimum variance portfolio			Mean variance portfolio		
Model	Rate	Std	Sharpe	Rate	Std	Sharpe
(1) Full covariance	-	-	-	-	-	-
(2) CAPM Empty	0.064	0.120	0.533	0.064	0.119	0.535
(3) CAPM Graph	0.074	0.107	0.688	0.075	0.107	0.700
(4) FF Empty	0.062	0.109	0.569	0.069	0.109	0.627
(5) FF Graph	0.070	0.105	0.661	0.072	0.106	0.678

vectors and variable selection. This analysis confirmed that the CAPM and FF model generally do well in explaining the variation of stock returns, but identifying relevant non-zero entries in the unexplained covariation are of real practical value: the resulting covariance matrix forecast has lower out-of-sample forecast errors, and the corresponding portfolios achieve lower level of realised risk in terms of variance and higher realised returns.

In addition to case studies, I have also provided a fully Bayesian framework of two-stage forecast of covariance matrices, a mechanism of graph evolution, and the use of sequential stochastic search for high-dimensional graphical model space.

In regards to the modeling of graphical structure through time, alternative approaches include the use of first-order Markov probabilities in which the graph ob-

tained at time t depends on which of the graphs obtained at time $t-1$, but not on what happened prior to $t-1$, and higher-order Markov probabilities that extend the dependence to graphs at time $t-2, t-2, \cdots$, etc. These alternatives require the learning of a higher-dimensional transition matrix between graphs. Even a sparse representation of the transition matrix, such as each graph only moves to its neighbours between two time points is limited in a sense that the sparse pattern would restrict the evolution of graphs between time.

The sequential stochastic search algorithm combines the sequential Monte Carlo idea and shotgun stochastic search algorithm. Exploration of a static model space to find high posterior probability graphs can be successfully carried out using direct search such as shotgun stochastic search method, certainly up to 100 vertices or so while traditional MCMC is competitive only for relatively small graphs (Jones et al., 2005). However, fast searching over a sequence of large model spaces is more challenging. This problem can be eased by noticing that from one step to the next I do not expect large changes in the mass of the distribution. Therefore, I could use the high probability graphs from the previous step as starting points to initiate a new search and rapidly traverse the graphical model space around these promising models.

5

Sparse seemingly unrelated regression modelling

5.1 Introduction

This chapter develops a sparse seemingly unrelated regression (SSUR) model with Gaussian errors; that is, a set of regressions in which both regression coefficients and error precision matrix have many zeros. Zeros in regression coefficients arise when each response possibly only depends on a subset of different predictors; zeros in a precision matrix arise when the error terms satisfy a set of conditional independence restrictions consistent with an underlying graphical model (Whittaker, 1990; Lauritzen, 1996). I study and propose a fully Bayesian analysis of the SSUR model, and provide effective methods for marginal likelihood computation using a specified subset of variables and a specified graphical model to structure the covariance matrix. This enables the simultaneous selections of variables and the covariance matrix as well as comparison of posterior inferences with respect to subsets and conditional independence structures.

Seemingly unrelated regression models (SUR) are frequently used in econometric, financial and sociological modelling (Zellner, 1962, 1971; Box & Tiao, 1973; Srivas-

tava & Giles, 1987). Computational issues of SUR relate to the basic challenges to the efficiency of statistical computation (simulation and optimisation) for SUR model parameters. Frequentist methods respond to the challenges by proposing efficient numerical and computational methods to optimise cost functions (Kontoghiorghes & Clarke, 1995; Foschi & Kontoghiorghes, 2002; Foschi et al., 2003). For more complex econometric applications, Markov chain Monte Carlo methods are used for Bayesian analyses of many variations of the SUR model (Chib & Greenberg, 1995; Smith & Kohn, 2000; Griffiths, 2001; Holmes et al., 2002). Other computational techniques for SUR inference include direct Monte Carlo style methods (Ando & Zellner, 2010). I build on prior work in non-sparse Bayesian SUR model analysis and develop MCMC methods for model fitting and computation for the sparse SUR model. I note that the synthesis of sparse regression models and sparse covariance matrix models has been considered by Cripps et al. (2005) and George et al. (2008), though with a different practical focus than the SUR model. These analysis indirectly model parsimonious covariance matrices on their re-parameterised forms; they address model uncertainty by computing posterior model probabilities without attempting to calculate the marginal likelihood by introducing a model indicator into a list of unknown parameters. To use these methods, one must specify all competing models, and carefully choose some tuning parameters to ensure that the chain mixes well in model space. I directly model the sparse inverse covariance matrix through use of conjugate priors; this leads to an efficient posterior sampling (Carvalho et al., 2007) with marginal likelihood calculated using Monte Carlo methods. Our examples show that these marginal likelihood approximations are adequate and useful in assessing alternative models.

In the context of dynamic SUR models, the graphical modelling of the covariance matrix of multivariate data appears in Carvalho & West (2007a,b); Wang & West (2009) and Wang et al. (2009). Our extension of the dynamic SSUR models gener-

alises this earlier work on the dynamic matrix-variate graphical model; I provide a fully Bayesian inference and model comparison related to both regression coefficient linear equality constraints and error intra-dependencies in the cross-sectional structure of the time series. This framework of dynamic SSUR models opens challenging methodology questions in core econometric modelling and computation.

5.2 Sparse seemingly unrelated regression modelling

5.2.1 Basic SUR models

To introduce the SSUR model, I begin with the usual SUR model. Consider p univariate dependent variables $y_{i,t}$ following individual regressions:

$$y_{i,t} = \mathbf{X}_{i,t}'\boldsymbol{\beta}_i + e_{i,t} \qquad t = 1, 2, \cdots, T, \tag{5.1}$$

where $\mathbf{X}_{i,t}$ is the n_i-vector of observations on n_i explanatory variables with possibly a constant term for individual i at time t, $\boldsymbol{\beta}_i = (\beta_{i1}, \cdots, \beta_{in_i})$ is a n_i-vector of unknown coefficients, and $e_{i,t}$ is a random error. Combine the model as follows:

(a) $\mathbf{y}_t = (y_{1,t}, \cdots, y_{p,t})'$, the $p \times 1$ observation vector;

(b)

$$\mathbf{X}_t = \text{diag}(\mathbf{X}_{1,t}, \mathbf{X}_{2,t}, \cdots, \mathbf{X}_{p,t}) = \begin{pmatrix} \mathbf{X}_{1,t} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{2,t} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_{p,t} \end{pmatrix},$$

the $n \times p$ matrix of observations on explanatory variables at time t with $n = \sum_{i=1}^{p} n_i$;

(c) $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \cdots, \boldsymbol{\beta}_p')'$, the $n-$vector of coefficients; and

(d) $\mathbf{e}_t = (e_{1,t}, \cdots, e_{p,t})'$, the $p-$vector of errors distributed as $N(\mathbf{0}, \mathbf{V})$.

Then the model is

$$\mathbf{y}_t \;=\; \mathbf{X}_t'\boldsymbol{\beta} + \mathbf{e}_t, \qquad\qquad t = 1, 2, \cdots, T. \tag{5.2}$$

The SUR model assumes the errors are contemporaneously correlated but not auto-correlated. In other words, if $\mathbf{e} = (\mathbf{e}_1', \cdots, \mathbf{e}_T')'$, the SUR model assumes $\text{cov}(\mathbf{e}) = \mathbf{I}_T \otimes \mathbf{V}$.

5.2.2 Variable selection in SUR

To introduce sparsity in SUR model parameters, I first extend the development of Bayesian variable selection for multiple regression models to the SUR models. Such an extension has been considered by Brown et al. (1998) in multivariate regression contexts. However, their model assumes all responses have the same predictors, and thus must generate a subset of predictors appropriate for all responses. The SUR model is broader and employs multivariate regression as a special case. The variable selection problem for SUR models arises when there is an unknown subset of $\mathbf{X}_i$ with regression coefficients so small in predicting y_i that it becomes preferable to ignore them. I let $\boldsymbol{\gamma} = (\gamma_{11}, \cdots, \gamma_{pm_p})$ index each of these 2^n possible subset choices, where $\gamma_{ij} = 0$ or 1 according to whether β_{ij} is small or large, respectively.

Prior distributions for parameters $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$ are taken as $\boldsymbol{\beta} = N(\mathbf{m}_0, \mathbf{H}_{\boldsymbol{\gamma}})$ where $\mathbf{m}_0$ and $\mathbf{H}_{\boldsymbol{\gamma}}$ must be specified. One convenient choice of $\mathbf{m}_0$ is a zero vector. The class of $\mathbf{H}_{\boldsymbol{\gamma}}$ may take the form $\mathbf{H}_{\boldsymbol{\gamma}} = \mathbf{D}_{\boldsymbol{\gamma}}\mathbf{R}_{\boldsymbol{\gamma}}\mathbf{D}_{\boldsymbol{\gamma}}$, following the univariate regression form of George & McCulloch (1993). Here $\mathbf{D}_{\boldsymbol{\gamma}}$ is a $n \times n$ diagonal matrix and $\mathbf{R}_{\boldsymbol{\gamma}}$ is a correlation matrix. The element of $\mathbf{D}_{\boldsymbol{\gamma}}$ corresponding to β_{ij} is τ_{ij0} when $\gamma_{ij} = 0$ and τ_{ij1} when $\gamma_{ij} = 1$. Particular considerations about τ_{ij0}, τ_{ij1} and $\mathbf{R}_{\boldsymbol{\gamma}}$ are discussed by George & McCulloch (1993, 1997). One convenient choice for $\mathbf{R}$ is $\mathbf{I}$, under which the elements of $\boldsymbol{\beta}$ are *a priori* independent. Another choice is a block diagonal matrix in which each block corresponds to the covariance matrix of the n_i-vector of $\boldsymbol{\beta}_i$.

5.2.3 Structured covariance matrix

The role of the covariance matrix $\mathbf{V}$ is one of the most important features of SUR models. The non-diagonality of the error covariance matrix usually entails that individual regression estimates using univariate linear model are sub-optimal; joint estimations of SUR that exploits the correlation between errors across equations may improve parameter estimation by the implied "data sharing". Motivation for our work relates to the increased dimension and complexity of the error covariance matrix. In this context, the covariance matrix must be understood in terms of structure and parsimony.

Substantial progress has been made on Bayesian covariance modelling by imposing structures. Structures are typically obtained by restricting the elements of a re-parameterisation for $\mathbf{V}$ (Daniels & Pourahmadi, 2002; Smith & Kohn, 2002; Chen & Dunson, 2003). The main issue in these approaches is that the induced prior on $\mathbf{V}$ depends on the ordering of elements. Instead, I now directly and parsimoniously model $\mathbf{V}$ by considering its restrictions induced by graphical model structuring.

I apply the theory and methodology of Gaussian graphical models of Section 1.2 to the error vector $\mathbf{e}_t$ where $\mathbf{e}_t \sim N(\mathbf{0}, \mathbf{V})$ with precision $\mathbf{\Lambda} = \mathbf{V}^{-1}$. Suppose $\mathbf{\Lambda}$ is constrained by a decomposable graph G. G then defines factorisations of SUR model densities. For notational clarity, I suppress subscript t. Now for any graph G, I have

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \mathbf{V}, G) = \prod_{P \in \mathcal{P}} p(\mathbf{y}_P \mid \mathbf{X}_P, \boldsymbol{\beta}_P, \mathbf{V}_P) \Big/ \prod_{S \in \mathcal{S}} p(\mathbf{y}_S \mid \mathbf{X}_S, \boldsymbol{\beta}_S, \mathbf{V}_S), \qquad (5.3)$$

where $\mathcal{P}$ is the set of complete prime components, or cliques, of G and $\mathcal{S}$ is the set of separators. For each subgraph $g \in \{\mathcal{P}, \mathcal{S}\}$, $\mathbf{y}_g$ is $|g|-$vector defined as $\mathbf{y}_g = \{y_i : i \in g\}'$, $\mathbf{X}_g$ is the corresponding design matrix defined as $\mathbf{X}_g = \text{diag}\{\mathbf{X}_i : i \in g\}$, $\boldsymbol{\beta}_g = \{\boldsymbol{\beta}_i' : i \in g\}'$, and $\mathbf{V}_g$ the corresponding sub-matrix of $\mathbf{V}$. Each term in equation (5.3) is multivariate normal, $\mathbf{y}_g \sim N(\mathbf{X}_g'\boldsymbol{\beta}_g, \mathbf{V}_g)$ with $\mathbf{\Lambda}_g = \mathbf{V}_g^{-1}$ having no

off-diagonal zeros.

For $\mathbf{V}$ and graph G, assume the prior is $\text{HIW}_G(b, \mathbf{D})$ with the following density function

$$p(\mathbf{V}) = \prod_{P \in \mathcal{P}} p(\mathbf{V}_P \mid b, \mathbf{D}_P) \Big/ \prod_{S \in \mathcal{S}} p(\mathbf{V}_S \mid b, \mathbf{D}_S),$$

where each component is an inverse Wishart density.

5.3 Posterior and marginal likelihood computation

5.3.1 Gibbs sampling on a given graph and variable index

Assume a SSUR model of Section 5.2.1 and priors of Section 5.2.2 and 5.2.3, and write $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_T)$ for the full set of data. It is easy to see that, on any specified graph G and index vector $\boldsymbol{\gamma}$, the posterior $p(\boldsymbol{\beta}, \mathbf{V} \mid \mathbf{Y})$ has conditionals:

$$(\mathbf{V} \mid \boldsymbol{\beta}, \mathbf{Y}) \;\sim\; \text{HIW}_G\{b + T, \mathbf{D} + \sum_{t=1}^{T} (\mathbf{y}_t - \mathbf{X}_t'\boldsymbol{\beta})(\mathbf{y}_t - \mathbf{X}_t'\boldsymbol{\beta})'\}, \tag{5.4}$$

$$(\boldsymbol{\beta} \mid \mathbf{V}, \mathbf{Y}) \;\sim\; \text{N}\{\mathbf{C}(\sum_{t=1}^{T} \mathbf{X}_t \mathbf{V}^{-1} \mathbf{y}_t + \mathbf{H}_{\boldsymbol{\gamma}}^{-1} \mathbf{m}_0), \mathbf{C}\}, \tag{5.5}$$

where $\mathbf{C} = \{\sum_{t=1}^{T} \mathbf{X}_t \mathbf{V}^{-1} \mathbf{X}_t' + \mathbf{H}_{\boldsymbol{\gamma}}^{-1}\}^{-1}$. These form the basis of an efficient Gibbs sampler to generate from the full posterior $p(\boldsymbol{\beta}, \mathbf{V} \mid \mathbf{Y})$. The Gibbs iterates involve sampling from the hyper-inverse Wishart and multivariate normal distribution. Simulation of the former is based on Carvalho et al. (2007).

5.3.2 Marginal likelihood approximation

Exploration of uncertainty about regression and graphical structures involves consideration of the marginal likelihood function over structures; namely

$$p(\mathbf{Y}) \equiv p(\mathbf{Y} \mid \boldsymbol{\gamma}, G) \;=\; \int p(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{V}) p(\boldsymbol{\beta}) p(\mathbf{V}) \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\mathbf{V}, \tag{5.6}$$

over $(\boldsymbol{\gamma}, G)$; the priors in the integrand depend on the index vectors and graphs although I drop that in the notation for clarity. The integral cannot be evaluated but I can generate useful approximations via use of Candidate's formula (Besag, 1989; Chib, 1995); there are other possible approximation methods that I will discuss in Section 6.3. Write $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \mathbf{V}\}$ for all parameters, and suppose that I can evaluate $p(\boldsymbol{\theta} \mid \mathbf{Y})$ for some subset of parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$; Candidate's formula gives the marginal likelihood via the identity $p(\mathbf{Y}) = p(\mathbf{Y}, \boldsymbol{\theta})/p(\boldsymbol{\theta} \mid \mathbf{Y})$. Applying this requires that I estimate components of the numerator or denominator. Choosing $\boldsymbol{\theta}$ to maximally exploit analytic integration is key, and different choices that integrate over different subsets of parameters will lead to different, parallel approximations of $p(\mathbf{Y})$ that can be compared. I use (A): $\boldsymbol{\theta} = \mathbf{V}$, and (B): $\boldsymbol{\theta} = \boldsymbol{\beta}$, giving two approximations based on marginalisation over desirably disjoint parameter subsets. Other choices might be considered though with less analytic tractability.

The marginal likelihood is theoretically given by each of

$$\text{(A): } p(\mathbf{Y}) \;=\; p(\mathbf{Y}, \mathbf{V})/p(\mathbf{V} \mid \mathbf{Y}) \text{ at } \textit{any} \text{ chosen value of } \boldsymbol{\theta} = \mathbf{V}, \tag{5.7}$$

and

$$\text{(B): } p(\mathbf{Y}) \;=\; p(\mathbf{Y}, \boldsymbol{\beta})/p(\boldsymbol{\beta} \mid \mathbf{Y}) \text{ at } \textit{any} \text{ chosen value of } \boldsymbol{\theta} = \boldsymbol{\beta}. \tag{5.8}$$

I estimate the components of these equations that have no closed form, then plug-in chosen values $\boldsymbol{\beta}, \mathbf{V}$ such as approximate posterior means, to provide two estimates of $p(\mathbf{Y})$. For (A), the numerator terms, $p(\mathbf{Y}, \mathbf{V}) = p(\mathbf{Y} \mid \mathbf{V})p(\mathbf{V})$, are each easily computed at any $\mathbf{V}$. The denominator term may be approximated by

$$p(\mathbf{V} \mid \mathbf{Y}) = \int p(\mathbf{V} \mid \mathbf{Y}, \boldsymbol{\beta}) p(\boldsymbol{\beta} \mid \mathbf{Y}) d\boldsymbol{\beta} \approx \frac{1}{M} \sum_{j=1}^{M} p(\mathbf{V} \mid \mathbf{Y}, \boldsymbol{\beta}^{(j)}),$$

where the sum is over posterior draws $\boldsymbol{\beta}^{(j)}$; this is easy to compute as it is a sum of the product of hyper-inverse Wishart densities. For (B), the numerator can be

analytically evaluated as $p(\mathbf{Y} \mid \boldsymbol{\beta})$. The density function in the denominator is approximated as

$$p(\boldsymbol{\beta} \mid \mathbf{Y}) = \int p(\boldsymbol{\beta} \mid \mathbf{V}, \mathbf{Y}) p(\mathbf{V} \mid \mathbf{Y}) \mathrm{d}\mathbf{V} \approx \frac{1}{M} \sum_{j=1}^{M} p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{V}^{(j)}),$$

where the sum over posterior draws $\mathbf{V}^{(j)}$ can be easily performed, with terms given by normal density evaluations.

5.3.3 Model space priors for variable selection and graphs

For the model space priors, I use beta-binomial priors for both variable space and graphical model space. The prior probability for a subset of variables is given by $p(\boldsymbol{\gamma} \mid w_{\boldsymbol{\gamma}}) = \prod_{i=1}^{p} w_{\boldsymbol{\gamma},i}^{k_{\boldsymbol{\gamma},i}} (1 - w_{\boldsymbol{\gamma},i})^{(n_i - k_{\boldsymbol{\gamma},i})}$ and each variable inclusion probability for regression i has a beta prior $w_{\boldsymbol{\gamma},i} \sim Be(a, b)$. This structure yields $p(\boldsymbol{\gamma}) = \prod_{i=1}^{p} B(a + k_{\boldsymbol{\gamma},i}, b + n_i - k_{\boldsymbol{\gamma},i})/B(a, b)$ where $B(a, b)$ is the beta function. The prior on graphical model space is $p(G \mid w_G) = w_G^{k_G}(1 - w_G)^{(m - k_G)}$ and $w_G \sim Be(c, d)$, for a graph G having k_G edges out of $m = 2^{p(p-1)/2}$ possible ones. The default uniform priors on the $w_{\boldsymbol{\gamma},i}$s and w_G imply a marginal prior

$$p(\boldsymbol{\gamma}) = \prod_{i=1}^{p} \frac{k_{\boldsymbol{\gamma},i}!(n_i - k_{\boldsymbol{\gamma},i})!}{(n_i + 1)n_i!} = \prod_{i=1}^{p} \frac{1}{(n_i + 1)} \binom{n_i}{k_{\boldsymbol{\gamma},i}}^{-1} \text{and}$$

$$p(G) = \frac{k_G!(m - k_G)!}{(m + 1)m!} = \frac{1}{(m + 1)} \binom{m}{k_G}^{-1}.$$

This choice of the model space prior is based on the consideration that the fully Bayesian priors have automatic adjustment for multiple testing as the numbers of possible variables and edges grow (Scott & Carvalho, 2008; Carvalho & Scott, 2009).

5.3.4 Direct Metropolis-Hastings-within-Gibbs algorithms

I now extend Markov chain Monte Carlo for variable selection (George & McCulloch, 1993; Geweke, 1996b; George & McCulloch, 1997; Madigan & York, 1995; Raftery et al., 1997; Brown et al., 1998) and multivariate graphical models (Giudici & Green, 1999; Jones et al., 2005) to learning on $(\boldsymbol{\gamma}, G)$ in the above SSUR analysis. This relies on the computation of the unnormalised posterior over graphs, $p(\boldsymbol{\gamma}, G \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\gamma}, G)p(\boldsymbol{\gamma}, G)$, for any specified model $(\boldsymbol{\gamma}, G)$. Jones et al. (2005) discuss performance of various stochastic search methods in single multivariate graphical models; for modest dimensions, they recommend simple local-move Metropolis-Hastings. Here, given a current pair $(\boldsymbol{\gamma}, G)$, I can apply local moves in G space based on the conditional posterior $p(G \mid \mathbf{Y}, \boldsymbol{\gamma})$, and vice-versa. A candidate G' is sampled from a proposal distribution $q(G'; G)$ and accepted with probability

$$\alpha = \min\{\ 1,\ p(G' \mid \mathbf{Y}, \boldsymbol{\gamma})q(G; G')/p(G \mid \mathbf{Y}, \boldsymbol{\gamma})q(G'; G)\ \};$$

our examples use the simple random add/delete edge move proposal of (Jones et al., 2005). I then couple this with a similar step using $p(\boldsymbol{\gamma} \mid \mathbf{Y}, G)$ at each iteration. This requires a Markov chain analysis on each variable and graph pair visited in order to evaluate marginal likelihood, so implying a substantial computational burden.

5.3.5 Indirect Metropolis-Hastings-within-Gibbs sampling algorithms

I can also simulate $G, \mathbf{V}, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ without eliminating the values of parameters $\boldsymbol{\beta}$ and $\mathbf{V}$ using the following Metropolis-Hastings-within-Gibbs sampler:

(a) $(G \mid \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto H(b, \mathbf{D}, G)/H\{b + T, \mathbf{D} + \sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{X}_t'\boldsymbol{\beta})(\mathbf{y}_t - \mathbf{X}_t'\boldsymbol{\beta})', G\}$ is sampled through local move Metropolis-Hastings algorithm;

(b) $(\mathbf{V} \mid \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\gamma}, G)$ is the same as in equation (5.4);

(c) $(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{V}, \boldsymbol{\gamma}, G)$ is the same as in equation (5.5);

(d) $(\boldsymbol{\gamma}_{ij} \mid \mathbf{Y}, \boldsymbol{\gamma}_{-ij}, \boldsymbol{\beta}, \mathbf{V}, G) \sim Bern\{u_{ij1}/(u_{ij0} + u_{ij1})\}$, where

$$\begin{aligned} u_{ij1} &= p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}_{-ij}, \boldsymbol{\gamma}_{ij} = 1) p(\boldsymbol{\gamma}_{-ij}, \boldsymbol{\gamma}_{ij} = 1) \text{ and} \\ u_{ij0} &= p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}_{-ij}, \boldsymbol{\gamma}_{ij} = 0) p(\boldsymbol{\gamma}_{-ij}, \boldsymbol{\gamma}_{ij} = 0). \end{aligned}$$

This indirect model search algorithm does not require Markov chain analysis for marginal likelihood approximations at each step, and hence it is much faster than the direct search. However, it is hard to assess whether such a stochastic search can avoid becoming stuck in a posterior mode in which case the use of the empirical frequency to represent posterior probability of a model is less informative than marginal likelihoods. Nevertheless, an initial run of this faster indirect sampling method can provide some useful $\boldsymbol{\gamma}$ and G values in order to start the direct model search based on the marginal likelihood.

5.4 Empirical exploration and comparison

5.4.1 A first simulated random sample

A sample of size $T = 60$ was drawn from a $p = 6$ SUR model given by

$$\begin{aligned} y_1 &= 1.3x_1 - 0.5x_3 + e_1, \\ y_2 &= 0.9x_1 - 0.3x_2 + 0.5x_3 + e_2, \\ y_3 &= x_1 + 0.5x_2 + 0.7x_3 + e_3, \\ y_4 &= 0.8x_4 - 0.6x_5 + e_4, \\ y_5 &= x_4 + 0.7x_5 + e_5, \\ y_6 &= 1.1x_4 - 0.6x_5 + e_6, \end{aligned}$$

where the x_i are draws from i.i.d. $N(0, 1)$, and the error covariance matrix is the autocovariance matrix of a stationary $AR(1)$ process with AR parameter 0.6 and innovation variance 1. To perform variable selection, I added six noisy variables to each regression equation, and so $n = 51$.

First consider an analysis on the true subset of variables and the graph under relatively vague priors with $\mathbf{m}_0 = \mathbf{0}, \tau_0 = 0.01, \tau_1 = 10, b = 3$ and $\mathbf{D} = 0.0001\mathbf{I}_6$. I have also tried other vague but proper hyper-parameter choices that lead to similar results. My current code is in Matlab. It takes about 20 seconds on a dual-cpu 2.4GHz desktop computer running CentOS 5.0 unix to generate 2000 MCMC iterates. Convergence is rapid and apparently fast-mixing in this simulation as well as in other simulated examples. Parallel checking for assessing the dual approximation of marginal likelihood, in Figure 5.1, shows an implementation check and illustrates the concordance of the two, parallel marginal likelihood estimates; these are very close and differ negligibly on the log probability scale. When compared with each other, method (A) of equation (5.7) generates more stable estimates across differing Monte Carlo sample sizes than method (B) of equation (5.8). This is probably because the posterior standard error of $\mathbf{V}$ is greater than that of $\boldsymbol{\beta}$.

Consider model uncertainty with model space priors in Section 5.3.3. I first ran an indirect stochastic search sampler with 10000 full iterations starting with the full model and $\boldsymbol{\beta} = \mathbf{0}$. The median probability model of $\boldsymbol{\gamma}$ is the true subset of variables excluding x_2 and x_3 in the regression of y_2, and the median probability model of G is the true underlying band diagonal graph. Repeat explorations suggest stability in the marginal likelihood estimation when smaller Monte Carlo sample sizes are used, and I use 2000 draws within each step of the model search. The direct add-delete Metropolis-within-Gibbs was run for 5000 iterations starting with the median probability model found in the initial indirect search. The most probable model visited, $(\hat{\boldsymbol{\gamma}}, \hat{G})$, is the true subset of variables and true underlying graph; these are local modes and also have the largest marginal likelihood. This model was first visited after 203 direct Markov chain steps.

I also estimated the variable and edge inclusion probability using the top 30 models identified. The variables and edges in the modes $(\hat{\boldsymbol{\gamma}}, \hat{G})$ generally have higher pos-

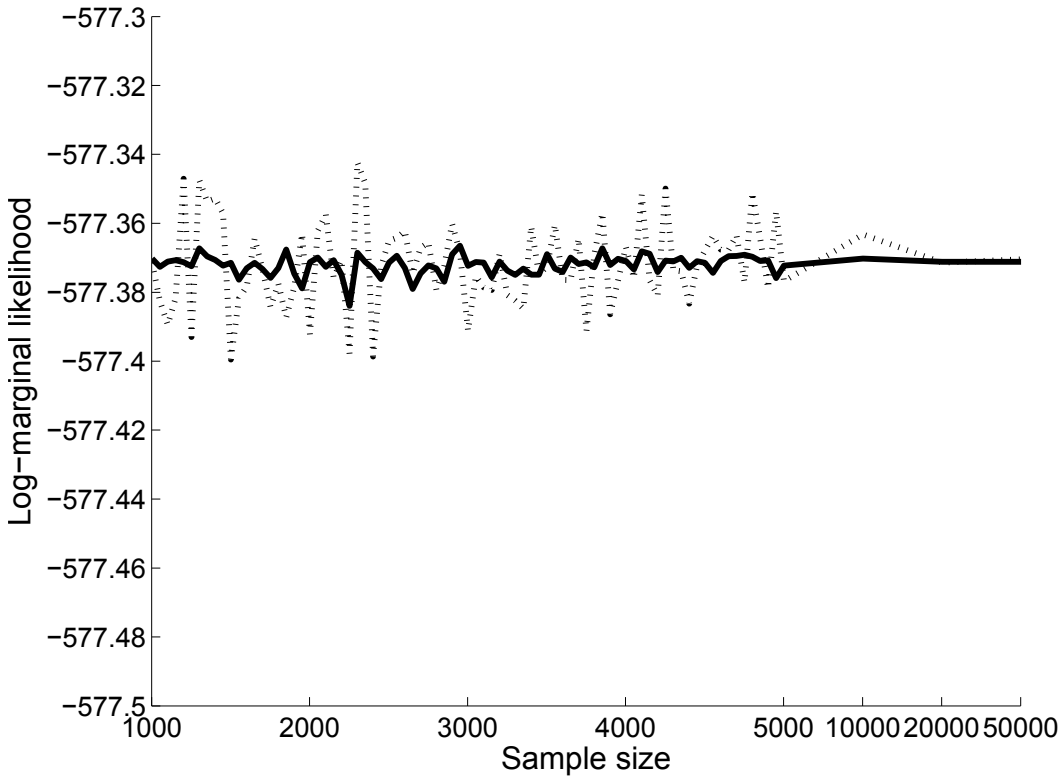

FIGURE 5.1: Log-marginal likelihood values on the true model in the simulation example of Section 5.4.1. The two estimates (solid line: method (A); dashed line: method (B)) of Section 5.3.2, were successively re-evaluated at differing simulation sample sizes. The plot confirms the concordance even at low samples sizes, and suggests very high accuracy in terms of differences on the log-likelihood scale.

terior inclusion probability than those not included; the lowest probability included variable and edge have probability 0.56 and 0.62 respectively, while the highest probability excluded variable and edge have probability 0.07 and 0.02 respectively. Thus, models discovered by highest posterior probability and by aggregating high probability models are not dramatically different. Further, the approximate posterior mean of the proportion of variables and edges, a measure of sparsity, are about $26\%, 29\%$ for $\boldsymbol{\gamma}, G$ respectively.

5.4.2 A second simulated example

The first simulated example demonstrated very good performance of SSUR modelling in identifying the underlying sparse structures. This second simulation study demonstrates that SSUR modelling can substantially improve parameter estimation. To compare the utility of SSUR with SUR, I compute the risk for $\boldsymbol{\beta}$ using the L_2 loss, $L(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \sum_{i,j}(\hat{\beta}_{i,j} - \beta_{i,j})^2$, and the risk for $\mathbf{V}$ using Stein's loss, $L(\hat{\mathbf{V}}, \mathbf{V}) = \text{tr}(\hat{\mathbf{V}}\mathbf{V}^{-1}) - \log|\hat{\mathbf{V}}\mathbf{V}^{-1}| - p$.

I conduct the second simulation in a scenario more suitable for econometric data. This model has $p = 15$ correlated explanatory variables. For each of T samples, I first generated $n_i = 10$ standard Gaussian features $\mathbf{X}_i$ with pairwise correlation 0.3 for each dependent variable y_i. I then randomly set the elements of the variable index $\boldsymbol{\gamma}$ to be 1 with probability 0.2, and 0 otherwise. The outcome $\mathbf{y}_t$ was generated according to a SUR model $\mathbf{y}_t = \mathbf{X}_t'\boldsymbol{\beta} + \mathbf{e}_t$ where $\mathbf{e}_t$ was generated from a multivariate Gaussian distribution with zero mean and covariance matrix $\mathbf{V}$. Each $\beta_{i,j}$ of the coefficient $\boldsymbol{\beta}$ was generated from a standard Gaussian distribution if $\gamma_{i,j} = 1$, and was set to be 0 if $\gamma_{i,j} = 0$. To specify $\mathbf{V}$, I fixed its correlation matrix $\mathbf{R}$ with inverse:

$$
\begin{pmatrix}
1.20 & -0.26 & \cdot & \cdot & \cdot & \cdot & \cdot & -0.40 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
-0.26 & 7.46 & \cdot & -0.23 & 0.46 & \cdot & -6.02 & \cdot & -1.37 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & 1.25 & \cdot & 0.56 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & -0.23 & \cdot & 6.99 & 0.76 & \cdot & -3.58 & \cdot & -0.98 & -0.64 & -2.40 & \cdot & \cdot & \cdot & \cdot \\
\cdot & 0.46 & 0.56 & 0.76 & 42.80 & \cdot & -37.71 & \cdot & \cdot & \cdot & -5.67 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & 1.90 & -0.69 & \cdot & \cdot & \cdot & -0.63 & \cdot & \cdot & \cdot & \cdot \\
\cdot & -6.02 & \cdot & -3.58 & -37.71 & -0.69 & 51.31 & \cdot & \cdot & \cdot & -3.98 & \cdot & \cdot & \cdot & \cdot \\
-0.40 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1.14 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & -1.37 & \cdot & -0.98 & \cdot & \cdot & \cdot & \cdot & 2.82 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & -0.64 & \cdot & \cdot & \cdot & \cdot & \cdot & 1.31 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & -2.40 & -5.67 & -0.63 & -3.98 & \cdot & \cdot & \cdot & 13.59 & -1.20 & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -1.20 & 1.80 & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 9.34 & -8.82 & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -8.82 & 9.34 & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1.00
\end{pmatrix},
$$

where $\cdot$ denotes zeros to highlight structure. I then chose the standard deviations $(v_1, \cdots, v_{15})$ so that the signal-to-noise ratio $Var\{E(y_i \mid \mathbf{X}_i)\}/Var(e_i)$ equalled 2.

For each simulated data set, I fit a SSUR model with the same hyper-parameters

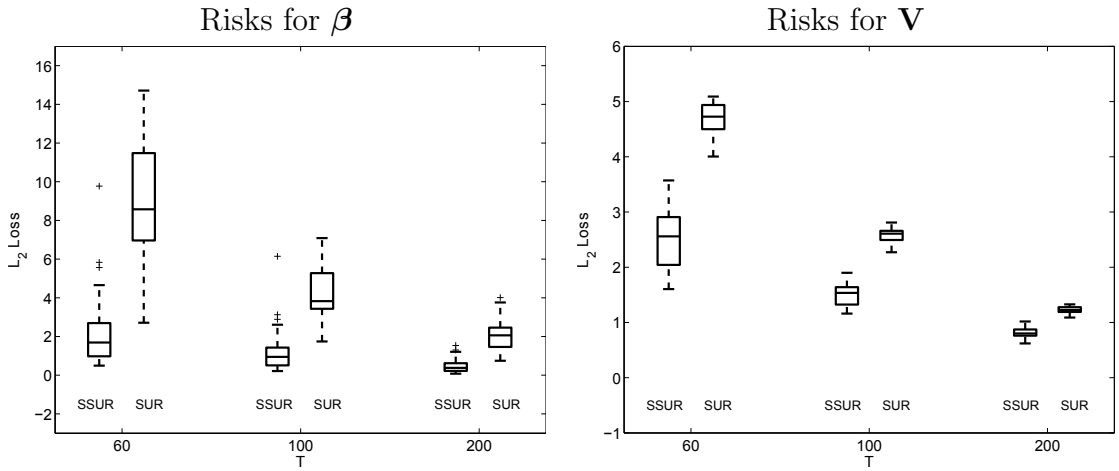

FIGURE 5.2: Estimation risk results for the second simulation study of Section 5.4.2. Shown are box plots of the risks over 100 simulations, for two estimators: SSUR and SUR, three different sample size: $T = 60, 100, 200$, and two sets of parameters: $\boldsymbol{\beta}$(left panel) and $\mathbf{V}$ (right panel).

as in the first simulation study and estimated the risks based on 10000 draws using the indirect algorithm of Section 5.3.5. For the SUR model, I used the same hyper-parameters of $\mathbf{m}_0, \tau_1, b$ and $\mathbf{D}$, and MC sample size 10000 based on the Gibbs algorithm in Section 5.3.1. I chose the sample size of $T = 60, 100, 200$ to see how well SSUR and SUR perform. Figure 5.2 shows boxplots of the risks achieved by the different estimators from SSUR and SUR models based on a total of 100 analysis of simulated data sets. It is clear that SSUR offers large gains over SUR in estimating both regression coefficients $\boldsymbol{\beta}$ and error covariance matrix $\mathbf{V}$. The gains are particularly significant when the sample size is small.

5.5 Example: Relations among stock returns, interest rates, real activity, and inflation

A p-dimensional vector autoregressive (VAR) process $\{\mathbf{y}_t\}$ with q lags can be formulated as SUR with dependent variables $\mathbf{y}_t$ and identical explanatory variables $\mathbf{X}_{i,t} = (1, \mathbf{y}'_{t-1} \cdots, \mathbf{y}'_{t-q})'$ for $i = 1, \cdots, p$. In this example, I use VAR models to investigate the relations and dynamic interactions among stock returns, interest rates,

real activity, and inflation in the postwar United States. The data are monthly real stock returns (SRE), real interest rates (IRE), industrial production growth (IPG) and inflation rates (INF). Real returns (SRE and IRE) are computed as nominal returns less the expected inflation rate. For comparison with the results in Lee (1992), the sample period for this study is from January 1947 to December 1987; the data appear in Figure 5.3.

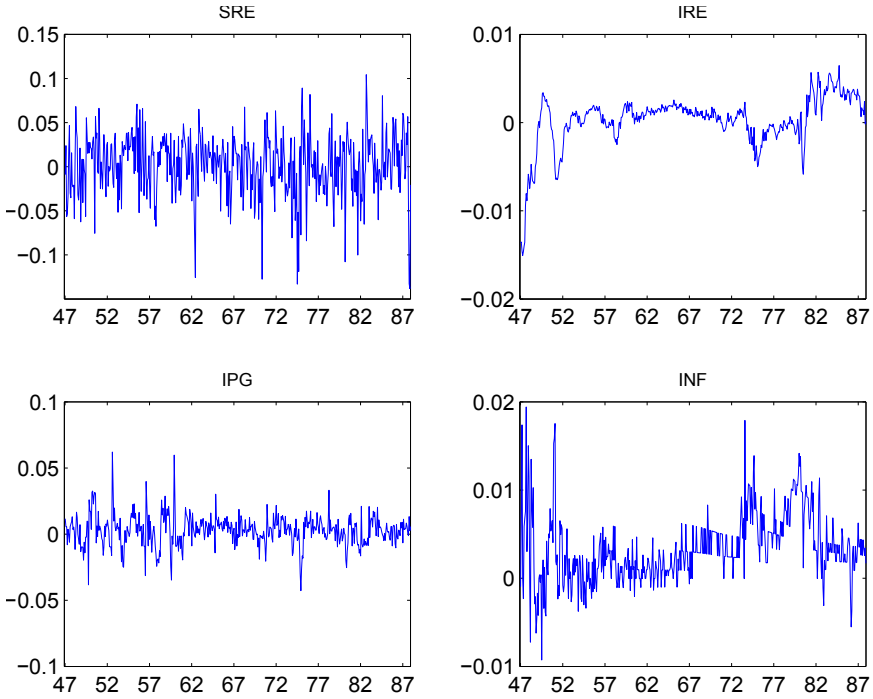

FIGURE 5.3: Monthly data on real stock returns (SRE), real interest rates (IRE), industrial production growth rates (IPG) and inflation rates (INF). The data set consists of 492 monthly rates for each of these four time series, over the period of 41 years: 01/1947 to 12/1987.

I choose a lag length of 6 months for the VAR model. Then, in month t, the dependent variables $\mathbf{y}_t$ and the explanatory variables $\mathbf{X_t}$ are given by $\mathbf{y}_t = (y_{1,t}, \cdots, y_{4,t})' = (\mathrm{SRE}_t, \mathrm{IRE}_t, \mathrm{IPG}_t, \mathrm{INF}_t)'$, and $\mathbf{X}_t = \mathrm{diag}(\mathbf{X}_{1,t}, \cdots, \mathbf{X}_{4,t})$ with

$$\begin{aligned} \mathbf{X}_{i,t} &= (1, \mathbf{y}'_{t-1} \cdots, \mathbf{y}'_{t-6})' \\ &= (1, \mathrm{SRE}_{t-1}, \mathrm{IRE}_{t-1}, \mathrm{IPG}_{t-1}, \mathrm{INF}_{t-1}, \cdots, \mathrm{SRE}_{t-6}, \mathrm{IRE}_{t-6}, \mathrm{IPG}_{t-6}, \mathrm{INF}_{t-6})', \end{aligned}$$

for $i = 1, \cdots, 4$. For variable selection, I use the default semiautomatic priors $\tau_{ij0} = 1/10\hat{\sigma}_{ij}$ and $\tau_{ij1} = 10\hat{\sigma}_{ij}$, where $\hat{\sigma}_{ij}$ is the standard error associated with the unconstrained generalised least squares estimate of β_{ij}. For covariance selection, I use the flat prior $b = 3$ and $\mathbf{D} = 0.0001\mathbf{I}_4$. The initial indirect stochastic search was run for 10000 steps, followed by a 10000 step run of direct stochastic search using marginal likelihood approximation based on 2000 Monte Carlo draws within each step. The marginal likelihood allows us to compute the exact relative probabilities using $p(\boldsymbol{\gamma}, G \mid \mathbf{Y})$. The relative probabilities of the 200 most probable models are displayed in order in Figure 5.4. This relative probability distribution is rather peaked, suggesting that a small subset of models are far more promising than others.

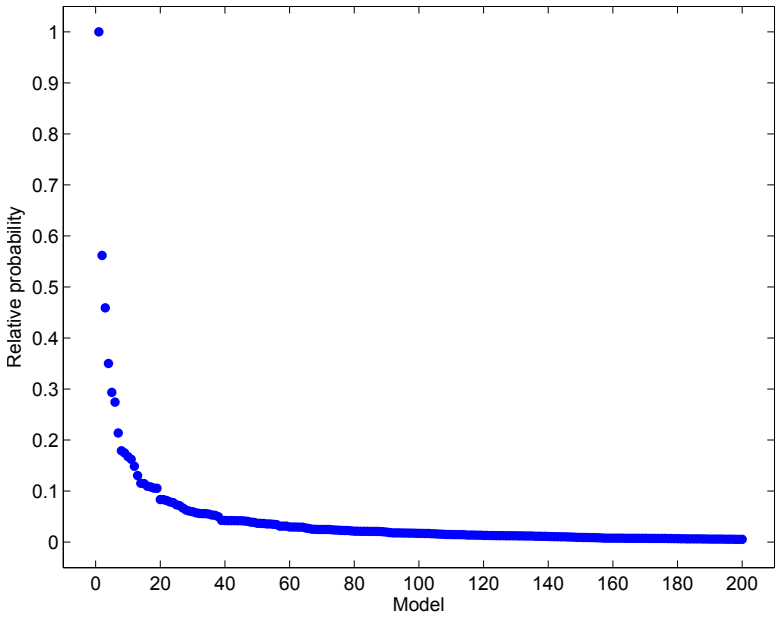

FIGURE 5.4: Relative posterior probabilities of the 200 most probable models. This peaked relative probability distribution implies that a small selection of high-probability models are far more better to be worked with than a grossly incorrect model on full graph and all of the predictors.

The following most probable models were reported with fitted VAR; the posterior

standard errors appear in parentheses:

$$
\begin{aligned}
y_{1,t} &= 0.248(0.044)y_{1,t-1} + e_1, \\
y_{2,t} &= 0.890(0.020)y_{2,t-1} - 0.076(0.011)y_{4,t-1} + e_2, \\
y_{3,t} &= 0.002(0.001) + 0.360(0.041)y_{3,t-1} + 0.047(0.014)y_{1,t-2} + 0.058(0.014)y_{3,t-3} + e_3, \\
y_{4,t} &= 0.001(0.0002) + 0.430(0.045)y_{4,t-1} + 0.184(0.045)y_{4,t-2} + 0.157(0.040)y_{4,t-5} + e_4.
\end{aligned}
$$

The estimated error adjacency matrix and covariance matrix are as follows:

$$
\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0.00113 & -4.86 \times 10^{-6} & 0 & -1.78 \times 10^{-6} \\ -4.86 \times 10^{-6} & 7.75 \times 10^{-7} & 0 & 2.84 \times 10^{-7} \\ 0 & 0 & 0.000105 & 0 \\ -1.78 \times 10^{-6} & 2.84 \times 10^{-7} & 0 & 8.35 \times 10^{-6} \end{pmatrix}.
$$

Real stock returns strongly signal positive first lag autocorrelation. Real interest rates appear to be highly positively autocorrelated, and are lead by inflation rates with a negative sign. Industrial product growth is positively autocorrelated, and is lead by real stock returns. Finally, inflation is also positively autocorrelated. The estimated error graph suggests that real stock returns are conditionally independent of inflation rates given real interest rates.

Table 5.1: Percentage of 24-month forecast error variance explained by innovations in each variable

	Variable Explained			
By Innovations in	SRE	IRE	IPG	INF
	(%)	(%)	(%)	(%)
SRE	97.3	0	8.54	0
IRE	2.66	68.2	0.234	0
IPG	0	0	91.2	0
INF	0.0334	31.8	0.00293	100

I further address the relations and interactions among these four variables by examining the percentage of 24-month orthogonal forecast error variance explained

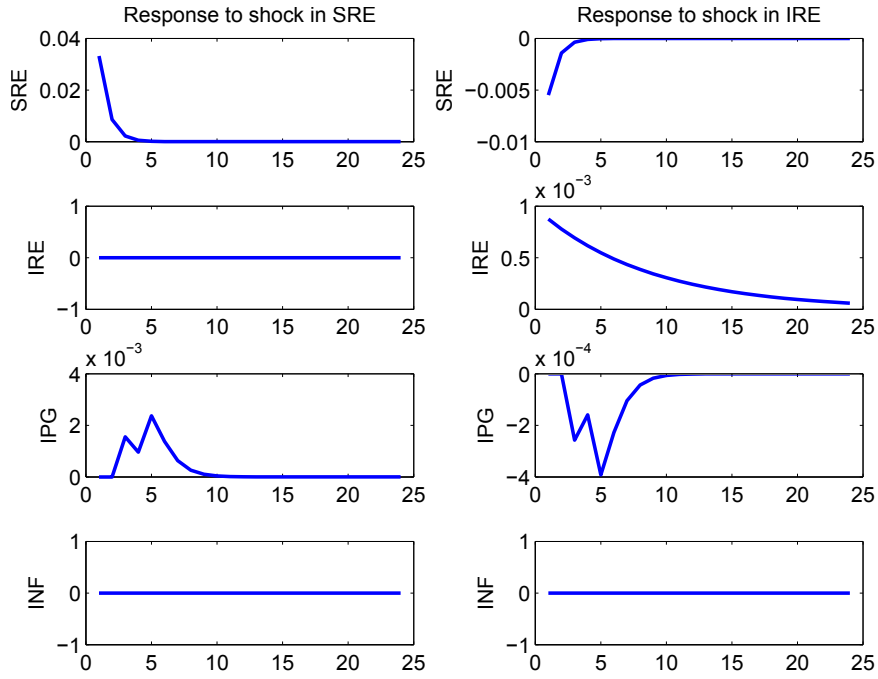

FIGURE 5.5: Estimated impulse response of each variable to shocks in real stock returns (left) and real interest rates (right).

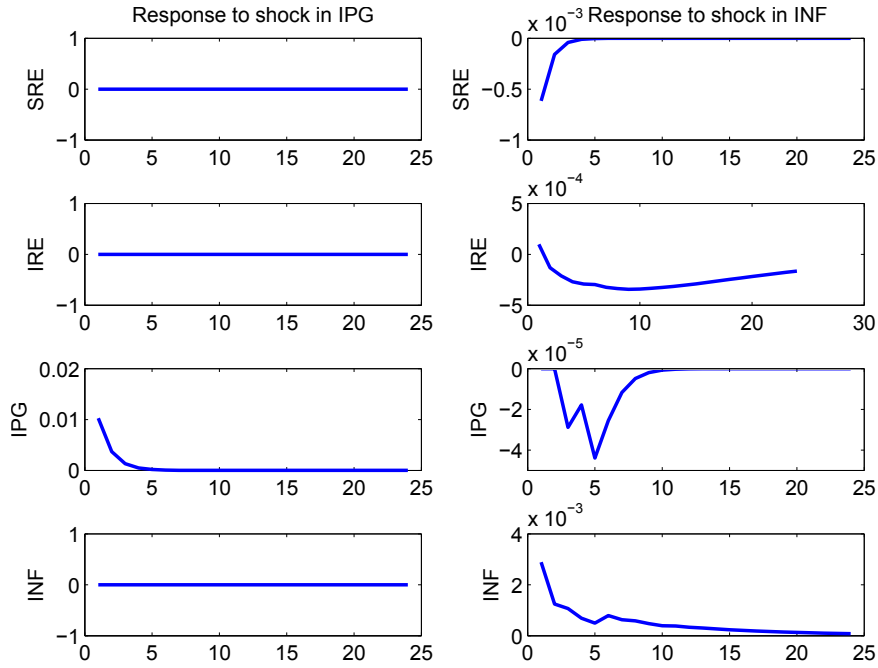

FIGURE 5.6: Estimated impulse response of each variable to shocks in real industrial production growth (left) and inflation rate (right).

by innovations in each variable as shown in Table 5.1 as well as the impulse response functions displayed in Figure 5.5 and 5.6.

(a) Real stock returns and real activity: Real stock returns appear to explain a substantial fraction (8.54%) of the variance in real activity, which responds positively to shocks in stock returns. Figure 5.5 shows that the response of industrial growth to shocks in real stock returns is significantly positive, peaks after five month and becomes negligible after ten months. This observation confirms the view that the stock market signals changes in real activity, and this correlation between stock returns and real activity is positive (Fama, 1981; Geske & Roll, 1983; Lee, 1992).

(b) Real stock returns and inflation: Real stock returns fail to Granger-cause inflation rates, since all of the coefficients on the lagged values of stock returns are zeros in the equation for inflation rates. Furthermore, the fourth column of Table 5.1 suggests that innovation in real stock returns explains none of the forecast error variance of inflation. This finding is generally compatible with the view that the negative observed relations between stock returns and inflation rates might be a proxy for other possible macroeconomic relations (Fama, 1981; Geske & Roll, 1983). In addition, Figure 5.5 suggests there is no signal of a consistent negative response of inflation to shocks in stock returns.

(c) Real interest rates and inflation: Contrary to the findings presented in Lee (1992), there is no indication that real interest rates Granger-cause inflation in this data set. Moreover, as is shown by column 4 in Table 5.1, innovations in real interest rates do not explain any of the forecast variance of inflation. However, inflation appears to explain a substantial fraction (31.8%) of forecast error variance of real interest rates.

(d) Inflation and real activity: Table 5.1 indicates that inflation only has negligible explanatory power (0.00003%) for real activity in the presence of real stock returns. Figure 5.6 shows that this weak relation between inflation and real activity is negative.

6

Extensions of sparse seemingly unrelated regression modelling

6.1 Mutual fund performance

6.1.1 Alpha and the SUR model

The historical performance of a mutual fund can be summarised by estimating its alpha. This term is defined as the intercept in a regression of the excess return of the fund on the excess return of one or more passive benchmarks. This is usually estimated by applying an ordinary least square analysis to the regression

$$y_{0,t} = \alpha_0 + \mathbf{x}'_{0,t}\boldsymbol{\beta}_0 + e_{0,t}, t = 1, 2, \cdots, T$$

where $y_{0,t}$ is the fund return at time t, $\mathbf{x}_t$ is a $k \times 1$ vector of benchmark returns at time t, and α_0 is the fund alpha. The choice of benchmarks is often guided by a pricing model, such as the capital asset pricing model (CAPM) (Sharpe, 1964; Lintner, 1965; Mossin, 1966) and the Fama-French three factor model (Fama & French, 1993). The recent work of Pástor & Stambaugh (2002) has explored the role of nonbenchmark passive assets in estimating a fund's alpha using a seemingly unrelated regression model. Suppose there are p nonbenchmark passive returns $y_{i,t}$

besides the k benchmark returns $x_{i,t}$. Then the SUR model used to estimate the mutual fund α_0 is written as

$$\begin{aligned} y_{0,t} &= \alpha_0 + \mathbf{x}_t'\boldsymbol{\beta}_0 + e_{0,t}, \\ y_{i,t} &= \alpha_i + \mathbf{x}_t'\boldsymbol{\beta}_i + e_{i,t}, i = 1, \cdots, p, \end{aligned} \tag{6.1}$$

where $e_t = (e_{0,t}, e_{1,t}, \cdots, e_{p,t})$ is correlated contemporaneously and not autocorrelated. The basic idea is that a more precise estimate of α_0 is provided through a more precise estimate of α_i when $e_{0,t}$ is correlated with the $e_{i,t}$ for all $i = 1, \cdots, p$. Note that many mutual funds have relatively short histories as compared with passive assets. Given the more accurate estimate of α_i $(i = 1, \cdots, p)$ computed from a longer sample period, the α_0 estimated from a SUR model is more precise than the α_0 estimated solely based on a single regression model.

6.1.2 Alpha and the SSUR model

Some interesting questions arise in evaluating mutual fund performance using SUR models. First, as is observed by Pástor & Stambaugh (2002), the assumption of pricing power of benchmark assets on nonbenchmark assets is critical in estimating a fund's α in a SUR model. In particular, if in each case the benchmark assets are assumed to have no pricing ability on the nonbenchmark assets, i.e. $\alpha_i \neq 0$ $(i = 1, \cdots, p)$, then the estimate of α_i from a longer sample period is more precise than the estimate of α_is from the same period of available history. Given the correlation between $e_{i,t}$ and $e_{0,t}$, the same can be said of the estimate of α_0 based on the SUR model relative to the estimate of α_0 from a single regression. Otherwise, if benchmark assets price other nonbenchmark assets, i.e. $\alpha_i = 0$ $(i = 1, \cdots, p)$, then the better performance of an estimate of α_0 based on the SUR model as compared to that based on a single regression is attributed to additional information about sampling error provided by the seemingly unrelated regressions of nonbench-

mark assets. Pástor & Stambaugh (2002) address the assumption of pricing power by separately applying SUR models to such situations. However, as is shown below, within the SSUR framework the uncertainty about the pricing power of benchmark assets on nonbenchmark assets can be incorporated naturally. The second interesting question concerns the strictness of the SUR model assumption, that is, returns are assumed to be contemporaneously correlated with all nonbenchmark returns given the benchmark returns. For certain types of managed funds, perhaps only the errors from a subset of nonbenchmark assets are relevant in explaining returns of the fund. Including too many correlated nonbenchmark assets to estimate alpha will mean a potentially high misspecification risk. Hence the possibility that a SUR model can account for the subset of nonbenchmark assets correlated with a fund is very compelling.

I also note that the history of a fund is very likely to be shorter than that of the passive assets. In order to extend the basic SSUR model to allow one equation to have fewer observations than the others, I re-parameterise the models in equation (6.1). Suppose returns on passive assets including benchmark or nonbenchmark assets are constructed for the period from 1 to T and a mutual fund only has a history from t_0 to T where $t_0 \geq 1$. Notice that $e_{0,t} = \sum_{i=1}^{p} e_{i,t}\theta_i + \tilde{e}_{0,t} = \sum_{i=1}^{p}(y_{i,t} - \alpha_i - \mathbf{x}_t'\boldsymbol{\beta}_i)\theta_i + \tilde{e}_{0,t}$ if the errors are correlated contemporaneously. Equation (6.1) can then be rewritten as

$$y_{0,t} = \tilde{\alpha}_0 + \mathbf{x}_t'\tilde{\boldsymbol{\beta}}_0 + \sum_{i=1}^{p} y_{i,t}\theta_i + \tilde{e}_{0,t}, \qquad t = t_1, \cdots, T, \tag{6.2}$$

$$y_{i,t} = \alpha_i + \mathbf{x}_t'\boldsymbol{\beta}_i + e_{i,t}, \qquad i = 1, \cdots, p,\ t = 1, \cdots, T, \tag{6.3}$$

where $\tilde{\alpha}_0 = \alpha_0 - \sum_{i=1}^{p} \alpha_i\theta_i$, $\tilde{\boldsymbol{\beta}}_0 = \boldsymbol{\beta}_0 - \sum_{i=1}^{p} \boldsymbol{\beta}_i\theta_i$ and $\tilde{e}_{0,t} \sim N(0, \tilde{\sigma}^2)$ is uncorrelated with the error vector $(e_{1,t}, \cdots, e_{p,t})$, which is distributed as $N(\mathbf{0}, \boldsymbol{\Sigma})$. I further assume throughout this section that the benchmark assets $\mathbf{x}_t$ are included in every possible

model. For equation (6.2), the models for different subsets of nonbenchmark assets may be represented by a vector of binary variables, $\boldsymbol{\gamma}_0 = (\gamma_{00}, \gamma_{01}, \cdots, \gamma_{0p})'$, where γ_{0j} is an indicator of the inclusion of intercept $\tilde{\alpha}_0$, when $j = 0$, or nonbenchmark asset y_j, when $j \geq 1$. For the p equations in equation (6.3), I index each of the possible benchmark assets' pricing abilities by $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)'$ where $\gamma_i = 0$ or 1 according to whether α_i is small or large, respectively. I use G to denote the graph underlying $\boldsymbol{\Sigma}$, which is the error covariance matrix of nonbenchmark passive assets.

Two interesting questions can now be addressed by incorporating model uncertainty regarding the choice of the triple $M = (\boldsymbol{\gamma}_0, \boldsymbol{\gamma}, G)$. Define

$$\boldsymbol{\Theta}_0 = (\tilde{\alpha}_0, \tilde{\boldsymbol{\beta}}_0, \theta_1, \cdots \theta_p, \tilde{\sigma}^2), \boldsymbol{\Theta}_1 = (\alpha_1, \cdots, \alpha_p, \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_p, \boldsymbol{\Sigma}), \mathbf{Y}_0 = (y_{0,t_1}, \cdots, y_{0,T}),$$

and

$$\mathbf{Y}_1 = (y_{1,1}, \cdots, y_{1,T}, \cdots, y_{p,1}, \cdots, y_{p,T}).$$

The likelihood function for $(\boldsymbol{\Theta}_0, \boldsymbol{\Theta}_1)$ can be factorised as

$$l(\boldsymbol{\Theta}_0, \boldsymbol{\Theta}_1) = p(\mathbf{Y}_0, \mathbf{Y}_1 \mid \boldsymbol{\Theta}_0, \boldsymbol{\Theta}_1) = p(\mathbf{Y}_0 \mid \mathbf{Y}_1, \boldsymbol{\Theta}_0) p(\mathbf{Y}_1, \boldsymbol{\Theta}_1).$$

Let $\mathcal{M} = \{M_k\}_k$ be the set of all possible triples M. For each model $M = M_k \in \mathcal{M}$, I assume the prior $p(\boldsymbol{\Theta}_0, \boldsymbol{\Theta}_1 \mid M_k) = p(\boldsymbol{\Theta}_0 \mid \boldsymbol{\gamma}_0) p(\boldsymbol{\Theta}_1 \mid \boldsymbol{\gamma}, G)$ where $p(\boldsymbol{\Theta}_0 \mid \boldsymbol{\gamma}_0)$ is the fully conjugate variable selection priors, and $p(\boldsymbol{\Theta}_1 \mid \boldsymbol{\gamma}, G)$ is the prior discussed in Sections 5.2.2 and 5.2.3. Coupling the likelihood and separable prior yields the full marginal likelihood of the data under model M_k:

$$\begin{aligned} p(\mathbf{Y} \mid M_k) &= \int p(\mathbf{Y} \mid \boldsymbol{\Theta}_0, \boldsymbol{\Theta}_1) p(\boldsymbol{\Theta}_0 \mid M_k) p(\boldsymbol{\Theta}_1 \mid M_k) \mathrm{d}\boldsymbol{\Theta}_0 \mathrm{d}\boldsymbol{\Theta}_1 \\ &= \int p(\mathbf{Y}_0 \mid \mathbf{Y}_1, \boldsymbol{\Theta}_0) p(\boldsymbol{\Theta}_0 \mid \boldsymbol{\gamma}_0) \mathrm{d}\boldsymbol{\Theta}_0 \int p(\mathbf{Y}_1 \mid \boldsymbol{\Theta}_1) p(\boldsymbol{\Theta}_1 \mid \boldsymbol{\gamma}, G) \mathrm{d}\boldsymbol{\Theta}_1, \end{aligned}$$

where the first integrand is available in closed form if using fully conjugate priors, and the second term is approximated by the Monte Carlo method as discussed in

the previous section. The posterior distribution over models M is given, at $M = M_k$ for each k, by

$$p(M_k \mid \mathbf{Y}) = \frac{p(M_k)p(\mathbf{Y} \mid M_k)}{\sum_k p(M_k)p(\mathbf{Y} \mid M_k)}.$$

If the fund's alpha, α_0, is of interest, I first transform back to α_0 using

$$\alpha_0 \quad = \quad \tilde{\alpha}_0\gamma_{00} + \sum_{i=1}^{p} \alpha_i\gamma_i\theta_i\gamma_{0i}.$$

Under model M_k, the posterior distribution for α_0 is a mixture over all models,

$$p(\alpha_0 \mid \mathbf{Y}) = \sum_k p(\alpha_0 \mid \mathbf{Y}, M_k)p(M_k \mid \mathbf{Y})$$

where $p(\alpha_0 \mid \mathbf{Y}, M_k)$ is the posterior distribution of α_0 under model M_k.

6.1.3 Vanguard managed funds

To evaluate the efficacy of the model, I applied it to a collection of 15 actively managed Vanguard mutual funds, using monthly returns through December 2008 available from the Center for Research in Security Prices (CRSP) mutual fund database. The names of the fund, the associated NASDAQ tickers and relevant inception dates are available in Table 6.1.

The set of benchmark and nonbenchmark assets consists of nine portfolios constructed passively. Monthly returns on these passive assets are available from January 1927 through December 2008. The sample period for any given mutual fund is a much shorter subset of this overall period. I specify the benchmark series as the excess market returns (MKT), and so the alpha is exclusively defined with respect to just MKT. The first two of nonbenchmark passive portfolios are the Fama-French factors, namely, SMB and HML, which are the payoffs on long-short spreads constructed by sorting stocks according to the market capitalisation and the book-to-market ratio.

The third nonbenchmark series, MOM, is the momentum factor. The remaining five nonbenchmark assets, denoted by IP1,IP2,IP3,IP4 and IP5, are the value-weighted returns for five industrial portfolios. All data and detailed descriptions of these nine series are publicly available at the data library of Professor Kenneth R. French [1].

For priors on $\mathbf{\Theta}_0$, I assume $\theta_i \sim N(0, \tilde{\sigma}^2\hat{\sigma}_i^2/100)$ if $\gamma_{0i} = 0$, and $\theta_i \sim N(0, 100\tilde{\sigma}^2\hat{\sigma}_i^2)$ if $\gamma_{0i} = 1$, where $\hat{\sigma}_i$ is the standard error of unconstrained OLS estimator $\hat{\theta}_i$, coupled with inverse gamma prior on $\tilde{\sigma}^2$, $\tilde{\sigma}^2 \sim IG(3/2, 9/2)$. For priors on $\mathbf{\Theta}_1$, I choose $\alpha_i \mid \gamma_i = 0 \sim N(0, 0.03^2)$, and $\alpha_i \mid \gamma_i = 1 \sim N(0, 1)$ for monthly α_i's and $i \geq 1$ in equation (6.3). This choice of hyperparameters is in line with the view that a yearly return of 0.36% in excess of the compensation for the risk borne may possibly be ignored; moreover,these excess yearly returns would be within 12%, The prior on the error covariance matrix was specified to provide weak prior knowledge, with $b = 3$ and $D = 0.0001\mathbf{I}_8$. Finally, I assume a uniform prior for $\boldsymbol{\gamma}_0$, and a model space prior as in Section 5.3.3 for $(\boldsymbol{\gamma}, G)$. In each of the 15 funds, the model space of $\boldsymbol{\gamma}_0$ has size $2^{10} = 1024$, which is small enough to be enumerated in a row. The model space of $(\boldsymbol{\gamma}, G)$ is of size $2^8 \times 2^{28}$. To explore this model space, the add-delete Metropolis-Hastings-within-Gibbs sampler was run for 20000 steps based on the marginal likelihood approximation from the 2000 Monte Carlo sample.

For models indexed by $(\boldsymbol{\gamma}, G)$, the most probable model is that $\gamma_i = 1$ for all $i = 1, \cdots, p$, with the residual graph pictured in Figure 6.1. This is also the median probability model. This modal model seems to suggest that the eight nonbenchmark assets are not all perfectly priced by the benchmark asset. The residual graph also indicates a great deal of conditional independencies among error terms.

For models indexed by $\boldsymbol{\gamma}_0$, Table 6.2 shows the inclusion probabilities for eight nonbenchmark assets for each of the 15 aggressive Vanguard funds. As can be seen, the errors between each one of the Vanguard managed funds and eight nonbench-

[1] see, http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

mark assets are contemporaneously correlated in different ways. The number of nonbenchmark asset regression equations that is related to a fund's regression equations varies from 0 (for US growth) and 8 (for Equity-Income). I note that a fund's contemporaneously dependencies on nonbenchmark assets seem to reflect a fund's portfolio composition. For example, the Capital Opportunity Fund seeks companies with long-term growth and has a 44.6% holding on the information technology sector as of May, 2008. The error of this fund is related to the error of nonbenchmark assets representing market capitalisation (SML) and high technology (IP3).

Table 6.3 reports the estimates of monthly α_0's within each fund based on the OLS, SUR and SSUR models for a five-year period, a ten-year period and the period since a fund's inception. The SSUR estimates are nontrivially different from their OLS and SUR counterparts. In particular, the α_0's tend towards zeros under the SSUR model. This is not surprising since the SSUR model assumes a positive probability for $\alpha_0 = 0$. One important issue in fund performance evaluation is whether the managed fund adds value beyond the standard passive benchmarks. I address this issue by computing the standard error of the three estimators of a fund's alpha. In Table 6.4 I examine the three standard errors. These standard errors reflect the precision of inferences about α_0. Two results are worth noting. First, the SUR standard errors are generally smaller than their OLS counterparts. This observation is compatible with that in Pástor & Stambaugh (2002). Second, with few exceptions, the SSUR model seems to reduce the standard error even more than the SUR model. Recall that the standard error of the SSUR estimates takes into account of structure uncertainty. The reduced standard errors seem to suggest that there is a great deal of sparsity within the SUR models and that identifying this sparsity can help provide more precise estimates of α_0's. Examining the results in Table 6.3 and 6.4 together, I find only a few funds have estimated α_0 that are two standard errors away from 0. This suggest that most of the 15 mutual funds do not generate excess returns

beyond the passive benchmark assets.

Table 6.5 reports the estimates of $\boldsymbol{\beta}_0$'s within each fund based on the OLS, SUR and SSUR models for a five-year period, a ten-year period and the period since a fund's inception. First, I note that quite different $\boldsymbol{\beta}_0$'s are generated when using OLS versus the two SUR-type models. Second, the difference in $\boldsymbol{\beta}_0$'s between the SUR and SSUR models is substantially less than that for the OLS model and SUR or SSUR models. This means that nonbenchmark assets play an important role in estimating the $\boldsymbol{\beta}_0$'s, and that imposing structures seems to affect $\boldsymbol{\beta}_0$'s less than adding nonbenchmark assets. The manner in which nonbenchmark assets provide information is illustrated most dramatically in the cases of higher-beta and lower-beta funds. For example, the Capital Opportunity Fund and Growth Equity funds have $\boldsymbol{\beta}_0$'s of 1.20 based on OLS model, while these figures decrease to 1 according to the two SUR models. The Dividend Growth and Equity-Income funds have $\boldsymbol{\beta}_0$'s of about 0.53 and 0.67 if estimated using OLS, while these figures are around 0.75 and 0.86 if estimated using the two SUR models. In Table 6.6 I examine the standard deviations of the three estimators. As evident in the table, these standard deviations are very similar.

6.2 Linear equality restrictions and dynamic SUR models

In this section, I consider two important extensions of the SSUR model given in Section 5.2. First, many economic applications of SUR models involve linear restrictions on the coefficients. For example, the same coefficients may appear in more than one equation, and so one may want to hypothesise that all equations have the same coefficient vector (Min & Zellner, 1993). In general, the main problem involves assessing the evidence in favour of a reduced model of the kind $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$, where $\mathbf{A}$ is a $r \times n$ matrix and $\mathbf{b}$ is a r-vector. Second, I allow the regression parameters to be time varying. In particular, the parameter vector $\boldsymbol{\beta}$ at time t is denoted by $\boldsymbol{\beta}_t$, and

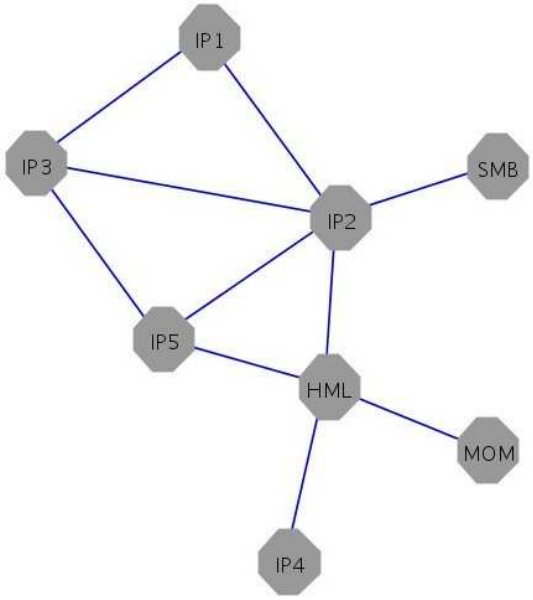

FIGURE 6.1: Highest log posterior graph of errors of nonbenchmark assets from the analysis of Vanguard funds

so the model is re-specified as

$$\begin{aligned} \mathbf{y}_t &= \mathbf{X}_t'\boldsymbol{\beta}_t + \mathbf{e}_t, \qquad \mathbf{e}_t \sim N(\mathbf{0}, \mathbf{V}), \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim N(\mathbf{0}, \mathbf{W}_t), \end{aligned} \tag{6.4}$$

with the initial prior $\boldsymbol{\beta}_0 \sim N(\mathbf{m}_0, \mathbf{C}_0)$. Carvalho & West (2007a,b); Wang & West (2009) used graphical model structuring for a covariance matrix in a class of dynamic SUR models that has been widely used to study financial time series (Quintana & West, 1987; Quintana, 1992; West & Harrison, 1997; Quintana et al., 2003). Their dynamic graphical model which leads to conjugate analysis requires that each univariate series $y_{i,t}$ must have the same predictors. In addition, $\mathbf{W}_t$ must be separable by a Kronecker product. Here I consider the general dynamic SUR model in equation (6.4) with two additional restrictions (a): $\mathbf{A}\boldsymbol{\beta}_t = \mathbf{b}$ for all t, and (b): $\mathbf{V}$ is constrained by one decomposable graph G.

Under the existence of the linear equality restriction (a), I reorder the elements

in $\boldsymbol{\beta}_t$ so that the restrictions can be written as

$$\mathbf{A}\boldsymbol{\beta}_t = (\mathbf{A}_1, \mathbf{A}_2) \begin{pmatrix} \boldsymbol{\beta}_{1,t} \\ \boldsymbol{\beta}_{2,t} \end{pmatrix} = \mathbf{b},$$

implying $\boldsymbol{\beta}_{1,t} = \mathbf{A}_1^{-1}(\mathbf{b} - \mathbf{A}_2\boldsymbol{\beta}_{2,t})$. Here $\mathbf{A}_1$ is $r \times r$ and nonsingular; $\mathbf{A}_2$ is $r \times (n-r)$; and $\boldsymbol{\beta}_{1,t}$ and $\boldsymbol{\beta}_{2,t}$ are r and $n-r$ sub-vectors of $\boldsymbol{\beta}_t$ respectively. Correspondingly, the observation equation in a dynamic SUR model can be written as

$$\mathbf{y}_t \;=\; \mathbf{X}_t'\boldsymbol{\beta}_t + \mathbf{e}_t = \mathbf{X}_{1,t}'\boldsymbol{\beta}_{1,t} + \mathbf{X}_{2,t}'\boldsymbol{\beta}_{2,t} + \mathbf{e}_t, \qquad \mathbf{e}_t \sim N(\mathbf{0}, \mathbf{V}),$$

or

$$\tilde{\mathbf{y}}_t \;=\; \tilde{\mathbf{X}}_t'\boldsymbol{\beta}_{2,t} + \mathbf{e}_t, \qquad \mathbf{e}_t \sim N(\mathbf{0}, \mathbf{V}), \tag{6.5}$$

where $\tilde{\mathbf{y}}_t = \mathbf{y}_t - \mathbf{X}_{1,t}'\mathbf{A}_1^{-1}\mathbf{b}$ and $\tilde{\mathbf{X}}_t' = \mathbf{X}_{2,t}' - \mathbf{X}_{1,t}'\mathbf{A}_1^{-1}\mathbf{A}_2$. Equation (6.5) is a SUR model without restrictions. Thus, any linear equality restricted dynamic SUR models can be transformed to unrestricted models. I only have to consider approaches to inference on the unrestricted dynamic SUR models of equation (6.4) as follows.

Suppose $\mathbf{W}_t$ is specified *a priori*. The inputs for the Gibbs sampler are as follows. Given $\mathbf{V}$, sampling the joint distribution of $(\boldsymbol{\beta}_{0:T} \mid \mathbf{Y}, \mathbf{V})$ is conducted using the forward filtering backward sampling algorithm detailed in West & Harrison (1997). The simulation of $\mathbf{V}$ for a specified graph is based on its full conditional distribution $(\mathbf{V} \mid \boldsymbol{\beta}_{0:T}, \mathbf{Y}) \sim \text{HIW}_G\{b+T, \mathbf{D}+\sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{X}_t'\boldsymbol{\beta}_t)(\mathbf{y}_t - \mathbf{X}_t'\boldsymbol{\beta}_t)'\}$. Furthermore, if Bayes factors to discriminate between models are of interests, these draws of $(\boldsymbol{\beta}_{0:T}, \mathbf{V})$ allow us to approximate the marginal likelihood in a similar manner to that in Section 3.1.

6.2.1 Example: Annual output growth rate data

An example concerns the choice of the pooled and the unpooled models for predicting annual output growth rates for industrialised countries. The data are taken from the IMF International Financial Statistics database for five countries, namely, Australia,

Canada, Japan,the UK and the USA. I fit the dynamic SUR model considered by Min & Zellner (1993) and Chib & Greenberg (1995). Here $y_{i,t}$ is et as the annual output growth rate for the ith country in the tth year, while

$$\mathbf{X}_{i,t} = (1, y_{i,t-1}, y_{i,t-2}, y_{i,t-3}, SR_{i,t-1}, SR_{i,t-2}, GM_{i,t-1}, MSR_{t-1})'$$

where $SR_{i,t}$ is the rate of growth of real stock prices, $GM_{i,t}$ is the rate of growth of real money, and MSR_t is the median of $SR_{i,t}$ in year t. Therefore, for each country i in year t, $\boldsymbol{\beta}_{i,t}$ is a vector of eight regression coefficients. The pooled model corresponds to the choice of $\boldsymbol{\beta}_t = \boldsymbol{\beta}_{1,t} = \cdots = \boldsymbol{\beta}_{5,t}$, while the unpooled model corresponds to the choice of $\boldsymbol{\beta}_t = (\boldsymbol{\beta}'_{1,t}, \cdots, \boldsymbol{\beta}'_{5,t})'$. Figure 6.2 displays the time series plots of y_t, SR_t and GM_t for each of the five countries.

It is worth noting that the Gibbs sampler for the dynamic SUR applies only when the sequence of state evolution variance matrices $\mathbf{W}_t$ is specified. This is different from the dynamic matrix-variate linear models in which $\mathbf{W}_t$ depends on $\mathbf{V}$ through a discount factor. In the general dynamic SUR model, if $\mathbf{W}_t$ depends on $\mathbf{V}$ through a discount factor, then such dependencies prevent the conditional distribution $(\mathbf{V} \mid \mathbf{Y}, \boldsymbol{\beta}_{0:T})$ from maintaining a tractable form. To identify a reasonable sequence of $\mathbf{W}_t$, the following strategy is used. First fit a static SUR model and estimate $\mathbf{V}$ using the Gibbs sampler from Section 5.3.1. Then specify $\mathbf{W}_t$ as $\mathbf{W}_t = (1-\delta)/\delta \mathbf{C}_{t-1}$, where $\mathbf{C}_{t-1}$ is the sequentially updated covariance matrix of $(\boldsymbol{\beta}_{t-1} \mid \mathbf{D}_{t-1})$ using an off-line estimated value for $\mathbf{V}$ from the static SUR model. In the analysis below, $\delta = 0.98$, and the four models represent different combinations of linear constraints and graphs G, where the linear constraint means either a pooled or an unpooled model and G is either a full graph or an empty graph. For each case, the Gibbs sampler was run 20 times each of which generated 50000 draws from the posterior distribution after discarding the first 5000 draws. The iterations began with the specification of values of $\mathbf{m}_0 = 0$, $\mathbf{C}_0 = 10\mathbf{I}_{40}$, $b = 3$ and $\mathbf{D} = 0.0001\mathbf{I}_5$. The results are summarised in

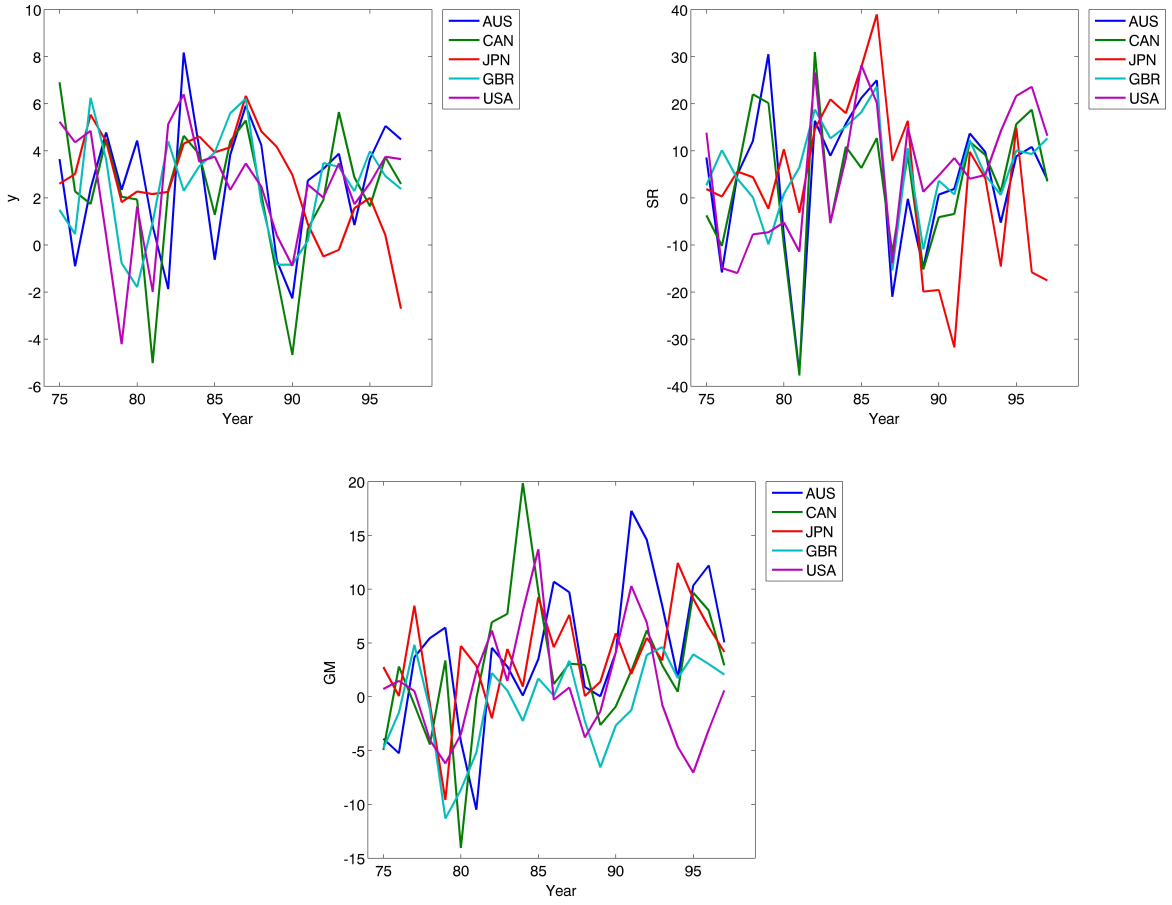

FIGURE 6.2: Time series plots of y_t as the annual output growth rate (upper left), SR_t as the rate of growth of real stock prices (upper right), and GM_t as the rate of growth of real money (bottom).

Table 6.7, where for each of four models, the log of the marginal likelihoods from two approximations are reported along with their numerical standard errors. Based on this table, it appears that the marginal likelihood is precisely estimated in all the fitted models. As expected, the more complex model has a larger numerical standard error associated with the estimation; for example, the numerical standard error of the unpooled model with a full graph is 147 times greater than that of the pooled model with an empty graph. These marginal likelihoods support the conclusion of Min & Zellner (1993); Chib & Greenberg (1995), who argue that a pooled model is

better than an unpooled model. Moreover, these marginal likelihoods indicate that the error covariance matrix may be diagonal.

6.3 Closing comments

SUR models are common in econometric studies. It is recognised that the conventional unconstrained SUR models may be over-parametrised. To remedy this problem, I have introduced a Bayesian analysis of the sparse seemingly unrelated regression (SSUR) model. The main innovations include inferences via Markov chain Monte Carlo simulations for specific constraints of regression coefficients and errors, evaluations of the marginal likelihoods of restrictions using coupled Candidate's formula approximations, and the extension of sparse modelling to dynamic SUR models.

Regarding the use of the conjugate hyper-inverse Wishart prior for the covariance matrix, Rajaratnam et al. (2008) provide theoretical support for the method used to estimate higher-dimensional covariance and precision matrices in Gaussian graphical models; in our models, this prior induces tractable and computationally accessible posteriors, leads to an adequate mixing of Markov chain simulations, and produces different approximations to marginal likelihoods of restrictions using the Candidate's formula.

Our use of Candidate's formula based on different approximations to marginal likelihoods is effective and efficient, as tested in a range of synthetic and real studies (Wang & West, 2009). There are other methods for approximately computing marginal likelihood, such as those proposed by Gelfand & Dey (1994) and Meng & Wong (1996). These methods often require the choice of some tuning functions. In contrast, the Candidate's formula is straightforward and easy to be implemented. These are also closely related non-Bayesian model selection criteria, including the well-known AIC, BIC, and extensions of them using information-theoretic ideas; these methods deserve further study. Here, I prefer the fully Bayesian approach that

enables the incorporation of prior information into the analysis.

The two real-world examples illustrate the important practical potential of the structured model. The first example investigates the causal relations and dynamic interactions among stock returns, interest rates, real activity, and inflation. Compared to previous studies, the SSUR model analysis identifies the important signals of dynamic relations among the variables without imposing *a priori* restrictions. The second example highlights the possibility of marginal likelihood estimation in a dynamic setting with general linear equality constraints.

A number of methodological issues remain. First, our examples are in modest dimensional problems where local move Metropolis-Hastings methods for the variable selection and graphical model components of the analysis can be expected to be effective. To scale to higher dimensions, alternatives computational strategies such as shotgun stochastic search over graphs (Dobra et al., 2004; Jones et al., 2005; Hans et al., 2007) become relevant. There is also potential for computationally faster approximations using expectation-maximisation style and variational methods (Jordan et al., 1999). Another issue involves the consideration of the structures in more complicated models with dynamic error covariance matrices such as those models examined in West & Harrison (1997) and Carvalho & West (2007a,b).

Table 6.1: Summary statistics of 15 Vanguard funds

		Annual Excessive Returns			
Name	Ticker	5 Year	10 Year	Since Inception	Inception Date
Cap Opp	VHCOX	−2.34	5.71	5.48	08/1995
Dividend Growth	VDIGX	−1.09	−3.07	2.19	05/1992
Equity-Income	VEIPX	−2.35	−1.30	4.55	03/1988
Explorer	VEXPX	−6.16	0.20	1.90	12/1967
Growth & Income	VQNPX	−6.09	−4.41	4.21	12/1986
Growth Equity	VGEQX	−8.13	−7.71	0.92	03/1992
Mid Cap Growth	VMGRX	−4.08	0.89	2.59	12/1997
Morgan Growth	VMRGX	−5.75	−4.23	3.35	12/1968
PRIMECAP	VPMCX	−1.38	0.52	8.01	11/1984
Selected Value	VASVX	−2.62	1.83	1.41	02/1996
Strategic Equity	VSEQX	−6.72	−10.95	6.32	08/1995
US Growth	VWUSX	−6.72	−10.95	6.32	01/1959
US Value	VUVLX	−5.28	−1.07	−1.07	06/2000
Windsor	VWNDX	−7.19	−1.57	4.70	10/1958
Windsor II	VWNFX	−3.68	−2.14	5.22	06/1985
Market	-	−4.43	−3.51	-	-

Table 6.2: Exact (to 2 decimal places) inclusion probabilities for 8 nonbenchmark assets for each of 15 aggressive Vanguard funds

Name	SML	HML	MOM	IP1	IP2	IP3	IP4	IP5
Cap Opp	1.00	0.04	0.03	0.05	0.26	0.94	0.05	0.04
Dividend Growth	0.10	1.00	0.04	0.06	0.74	0.09	0.34	0.05
Equity-Income	1.00	1.00	0.99	0.97	1.00	0.92	1.00	1.00
Explorer	1.00	1.00	0.25	0.06	0.36	0.15	0.18	0.21
Growth & Income	1.00	0.29	0.12	1.00	0.18	0.61	0.35	0.63
Growth Equity	0.07	1.00	1.00	0.82	0.06	1.00	0.07	0.13
Mid Cap Growth	1.00	0.60	1.00	0.35	0.04	0.14	0.03	0.04
Morgan Growth	1.00	1.00	0.49	1.00	0.87	1.00	0.22	0.81
PRIMECAP	1.00	0.15	0.04	0.60	0.58	1.00	0.31	0.32
Selected Value	0.99	0.94	1.00	0.93	1.00	0.06	0.22	0.48
Strategic Equity	1.00	0.85	0.06	0.28	1.00	0.08	0.08	0.98
US Growth	0.02	0.14	0.02	0.02	0.03	0.02	0.12	0.05
US Value	0.07	1.00	0.17	0.45	0.08	0.05	0.21	0.15
Windsor	0.19	1.00	1.00	0.89	1.00	0.99	0.07	1.00
Windsor II	1.00	1.00	1.00	0.98	1.00	0.10	1.00	1.00

Table 6.3: Estimated monthly α's from each of the three models: The least square estimates from the OLS, and the posterior mean estimates from the SUR and SSUR models. An asterisk symbol (*) flags an estimated α_0 that is two standard errors away from 0

	5 year			10 year			Since Inception		
Name	OLS	SUR	SSUR	OLS	SUR	SSUR	OLS	SUR	SSUR
Cap Opp	0.27	0.22	0.07	$*0.88$	$*0.81$	$*0.56$	0.34	0.33	0.18
Dividend Growth	0.16	$*0.31$	0.08	-0.08	-0.20	-0.02	0.05	-0.11	0.13
Equity-Income	0.06	$*0.31$	0.20	0.07	0.09	0.15	0.14	-0.01	-0.01
Explorer	-0.03	-0.03	0.06	0.43	0.08	0.16	-0.05	-0.16	0.03
Growth & Income	-0.16	-0.11	-0.04	-0.10	-0.01	0.05	0.02	-0.01	0.02
Growth Equity	-0.25	-0.09	-0.04	-0.20	-0.12	0.00	-0.20	-0.13	-0.02
Mid Cap Growth	0.11	0.14	0.02	0.61	0.38	0.23	0.55	0.43	0.25
Morgan Growth	-0.07	-0.09	-0.01	-0.01	0.00	0.05	0.04	-0.03	0.03
PRIMECAP	0.24	0.22	0.06	$*0.36$	$*0.34$	0.15	0.23	0.18	$*0.21$
Selected Value	0.14	-0.01	-0.11	0.39	0.17	0.17	0.09	-0.11	0.15
Strategic Equity	-0.01	-0.23	0.00	0.34	0.10	$*0.23$	0.14	0.00	0.10
US Growth	-0.20	0.03	0.02	$*-0.53$	-0.32	-0.17	0.31	0.29	0.01
US Value	-0.10	-0.11	0.00	0.31	0.20	$*0.17$	0.31	0.20	$*0.17$
Windsor	-0.23	-0.25	-0.19	0.15	0.06	0.14	0.14	-0.12	-0.11
Windsor II	0.02	0.07	0.00	0.04	-0.02	0.02	0.13	-0.03	-0.03

Table 6.4: Standard errors of each of the three estimates of monthly α's.

	5 year			10 year			Since Inception		
Name	OLS	SUR	SSUR	OLS	SUR	SSUR	OLS	SUR	SSUR
Cap Opp	0.23	0.21	0.18	0.27	0.21	0.27	0.26	0.22	0.11
Dividend Growth	0.12	0.12	0.08	0.25	0.19	0.28	0.18	0.15	0.14
Equity-Income	0.14	0.12	0.11	0.22	0.10	0.12	0.12	0.08	0.08
Explorer	0.19	0.11	0.10	0.28	0.15	0.10	0.14	0.12	0.10
Growth & Income	0.09	0.09	0.10	0.10	0.07	0.05	0.06	0.04	0.04
Growth Equity	0.22	0.15	0.16	0.24	0.15	0.12	0.16	0.11	0.12
Mid Cap Growth	0.21	0.18	0.15	0.41	0.27	0.14	0.38	0.25	0.14
Morgan Growth	0.12	0.10	0.07	0.10	0.09	0.05	0.07	0.07	0.08
PRIMECAP	0.14	0.15	0.12	0.18	0.16	0.12	0.12	0.11	0.09
Selected Value	0.20	0.20	0.14	0.32	0.20	0.13	0.28	0.19	0.23
Strategic Equity	0.18	0.15	0.09	0.20	0.14	0.09	0.17	0.12	0.17
US Growth	0.19	0.15	0.11	0.21	0.18	0.13	0.26	0.28	0.09
US Value	0.11	0.12	0.08	0.17	0.12	0.07	0.17	0.12	0.07
Windsor	0.14	0.14	0.10	0.24	0.14	0.09	0.09	0.09	0.09
Windsor II	0.15	0.12	0.10	0.25	0.11	0.15	0.12	0.08	0.08

Table 6.5: Estimated $\boldsymbol{\beta}$'s from each of the three models: The least square estimates from the OLS, and the posterior mean estimates from the SUR and SSUR models.

	5 year			10 year			Since Inception		
Name	OLS	SUR	SSUR	OLS	SUR	SSUR	OLS	SUR	SSUR
Cap Opp	1.17	1.03	1.09	1.23	1.04	1.02	1.20	1.05	1.07
Dividend Growth	0.72	0.80	0.75	0.56	0.83	0.80	0.53	0.74	0.75
Equity-Income	0.72	0.83	0.80	0.60	0.85	0.86	0.67	0.86	0.86
Explorer	1.23	1.20	1.19	1.20	1.24	1.23	1.11	1.02	1.04
Growth & Income	0.96	1.00	0.99	0.93	0.93	0.95	0.98	0.97	0.97
Growth Equity	1.11	1.00	1.02	1.29	1.00	1.00	1.25	0.98	0.96
Mid Cap Growth	1.15	1.05	1.07	1.37	1.20	1.25	1.36	1.16	1.23
Morgan Growth	1.07	0.99	1.02	1.10	0.98	0.99	1.06	0.96	0.95
PRIMECAP	0.96	0.93	0.95	1.04	0.94	0.93	1.07	0.97	1.00
Selected Value	0.96	1.00	0.98	0.71	1.02	1.01	0.77	1.10	1.12
Strategic Equity	1.23	1.21	1.24	0.95	1.17	1.17	0.98	1.18	1.19
US Growth	0.98	0.90	0.90	1.19	0.97	0.98	1.09	0.96	1.08
US Value	0.93	0.97	0.95	0.82	1.03	1.01	0.82	1.03	1.01
Windsor	0.99	1.00	1.01	0.87	1.07	1.08	0.93	1.05	1.05
Windsor II	0.89	0.98	0.96	0.69	1.00	0.99	0.81	1.02	1.02

Table 6.6: Standard errors of each of the three estimates of $\boldsymbol{\beta}$'s.

	5 year			10 year			Since Inception		
Name	OLS	SUR	SSUR	OLS	SUR	SSUR	OLS	SUR	SSUR
Cap Opp	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.07
Dividend Growth	0.03	0.03	0.03	0.05	0.05	0.06	0.04	0.05	0.04
Equity-Income	0.03	0.03	0.04	0.05	0.03	0.03	0.03	0.02	0.02
Explorer	0.05	0.03	0.03	0.06	0.04	0.04	0.03	0.03	0.03
Growth & Income	0.02	0.03	0.02	0.02	0.02	0.02	0.01	0.01	0.01
Growth Equity	0.06	0.04	0.04	0.05	0.04	0.04	0.04	0.03	0.03
Mid Cap Growth	0.05	0.05	0.05	0.09	0.08	0.08	0.08	0.07	0.07
Morgan Growth	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02
PRIMECAP	0.03	0.04	0.04	0.04	0.05	0.05	0.03	0.03	0.03
Selected Value	0.05	0.06	0.05	0.07	0.05	0.06	0.06	0.05	0.06
Strategic Equity	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.03
US Growth	0.05	0.04	0.04	0.05	0.05	0.05	0.06	0.08	0.07
US Value	0.03	0.04	0.03	0.04	0.03	0.03	0.04	0.03	0.03
Windsor	0.04	0.04	0.03	0.05	0.04	0.04	0.02	0.02	0.02
Windsor II	0.04	0.03	0.03	0.05	0.03	0.03	0.03	0.02	0.02

Table 6.7: Summary of results for output growth rate data. Each of the two approximations were run for 20 times. The mean and the numeric standard error of log marginal likelihoods from these 20 runs are reported for each of the approximations.

Model fitted	Log(marginal)		Numerical SE	
	(A)	(B)	(A)	(B)
Unpooled, Full graph	81.31	81.27	0.28	0.23
Unpooled, Empty graph	89.89	89.88	0.0047	0.083
Pooled, Full graph	171.25	171.25	0.0069	0.0078
Pooled, Empty graph	193.80	193.80	0.0019	0.0033

7

Concluding remarks and future research

My thesis has considered a number of developments of Bayesian multivariate analysis in dynamic models and graphical models. The main contributions are the development of novel sparse modelling and efficient computational techniques for several different types of important models: Matrix-variate time series, dynamic covariance models, and multivariate regressions. I applied these methods to real and simulated finance and econometrics data sets. These methods are also applicable to other areas such as biomedical sciences, spatial sciences and social sciences, where large data sets are routinely generated and often require sparse modelling. In addition to methodology produced by this thesis, the work has opened up abundant research opportunities in Bayesian graphical models and time series analysis. I will now briefly describe some future directions that are related to these two areas.

7.1 Future work on matrix normal graphical models

The application of matrix normal graphical models of Section 2 and 3 relies on the assumption of separability of the covariance matrix of the vector vec($\mathbf{Y}$) formed by stacking the columns of $\mathbf{Y}$ into a single column vector. However, these separable

models are not always appropriate, and formal tests are needed for the hypothesis of separability of covariance matrices. In a frequentist framework, likelihood ratio tests based on asymptotic theory have been considered (Lu & Zimmerman, 2005; Mitchell et al., 2005) in the context of iid samples. In a Bayesian framework, no formal methodology has been developed for the problem of testing separability of covariance matrices. Moreover, there has been no work on examining the separability of covariance matrices that are structured by graphs.

The formal definition of testing for separability can be described as follows. I define M_r^+ to be the space of all positive definite matrices of order r. Let $M_p^+ \otimes M_q^+ \subset M_{pq}^+$. To test the hypothesis $H_0 : \mathbf{W} = \mathbf{V} \otimes \mathbf{U} \in M_p^+ \otimes M_q^+$ versus $H_1 : \mathbf{W} \in M_{pq}^+$, I put priors π_0 on $M_p^+ \otimes M_q^+$ and π on M_{pq}^+. If I observe data $\mathbf{Y} = \mathbf{Y}_1, \cdots, \mathbf{Y}_n$, the impact of data on model uncertainty can be isolated in the corresponding Bayes factor:

$$B_{01} = \frac{\int_{M_p^+ \otimes M_q^+} f(\mathbf{Y} \mid \mathbf{W})\pi_0(\mathbf{W})\mathrm{d}\mathbf{W}}{\int_{M_{pq}^+} f(\mathbf{Y} \mid \mathbf{W})\pi(\mathbf{W})\mathrm{d}\mathbf{W}}. \tag{7.1}$$

An important consideration for constructing a prior on the restricted parameter space is whether or not the priors π and π_0 are "compatible" (Dawid & Lauritzen, 2000). I want Bayes factor to truly respond to the data, rather than merely reflecting prior prejudices.

Suppose I adopt the projection approach of McCulloch & Rossi (1992). Specifically, I have associate each distribution π on M_{pq}^+ with a corresponding distribution π_0 on $M_p^+ \otimes M_q^+$. Such a specification can be achieved by an appropriate mapping $h : M_{pq}^+ \to M_p^+ \otimes M_q^+$ such that $\pi_0 = h(\pi)$. Thus, the numerator of the Bayes factor (7.1) can then be expressed as an unrestricted integral using projection function h:

$$\int_{M_p^+ \otimes M_q^+} f(\mathbf{Y} \mid \mathbf{W})\pi_0(\mathbf{W})\mathrm{d}\mathbf{W} = \int_{M_{pq}^+} f(\mathbf{Y} \mid h(\mathbf{W}))\pi(\mathbf{W})\mathrm{d}\mathbf{W}.$$

One way to define a projection h is based on minimising the discrepancy of projection. I could define a discrepancy function $D(\mathbf{W}, \mathbf{V} \otimes \mathbf{U})$ and choose $h(\mathbf{W})$ so as to minimise the discrepancy:

$$h(\mathbf{W}) = \arg\min D_{\mathbf{V} \in M_p^+, \mathbf{U} \in M_q^+}(\mathbf{W}, \mathbf{V} \otimes \mathbf{U}).$$

A popular discrepancy function is the Kullback-Leibler divergence:

$$D_{KL} = E_{f(\mathbf{Y}|\mathbf{W})}[\log\{f(\mathbf{Y} \mid \mathbf{W})/f(\mathbf{Y} \mid \mathbf{U}, \mathbf{V})\}].$$

In the multivariate Gaussian case, the Kullback-Leibler divergence between $N(\mathbf{0}, \mathbf{W})$ and $N(\mathbf{0}, \mathbf{\Sigma})$ is $D_{KL}(\mathbf{W}, \mathbf{\Sigma}) = 1/2\{\mathrm{tr}(\mathbf{W}\mathbf{\Sigma}^{-1} - \mathbf{I}) - \log\det(\mathbf{W}\mathbf{\Sigma}^{-1})\}$, and for $\mathbf{\Sigma}$ restricted to being the Kronecker product of $\mathbf{U}$ and $\mathbf{V}$, we therefore have

$$D_{KL}(\mathbf{W}, \mathbf{V} \otimes \mathbf{U}) = 1/2[\mathrm{tr}\{\mathbf{W}(\mathbf{V}^{-1} \otimes \mathbf{U}^{-1}) - \mathbf{I}\} - \log\det\{\mathbf{W}(\mathbf{V}^{-1} \otimes \mathbf{U}^{-1})\}]. \quad (7.2)$$

This can be minimised in terms of $\mathbf{U}, \mathbf{V}$ to obtain the map from each unrestricted variance matrix $\mathbf{W}$ to separated matrices $(\mathbf{U}, \mathbf{V})$.

The next theorem provides a map $h : M_{pq}^+ \to M_p^+ \otimes M_q^+$.

Theorem 3. *The minimiser of Kullback-Leibler divergence in equation (7.2) exists. If* $\mathbf{V}$ *satisfies the constraint that* $v_{11} = 1$*, such minimiser is uniquely defined by the solution* $(\mathbf{U}, \mathbf{V})$ *of the following equations:*

$$\mathbf{U} = \frac{1}{p} \sum_{1 \le i,j \le p} \lambda_{ij} \mathbf{W}_{i,j}, \quad (7.3)$$

$$\mathbf{V} = \frac{1}{q} \sum_{1 \le i,j \le q} \omega_{ij} \tilde{\mathbf{W}}_{i,j}, \quad (7.4)$$

where $\mathbf{W} = (\mathbf{W}_{ij})$*, the* $pq \times pq$ *unrestricted covariance matrix comprised of* $p \times p$ *blocks* $\mathbf{W}_{ij}$ *of dimension* $q \times q$*, and* $\tilde{\mathbf{W}} = \mathbf{K}_{qp}\mathbf{W}\mathbf{K}_{pq} = (\tilde{\mathbf{W}}_{ij})$*, the* $pq \times pq$ *permutated*

unrestricted covariance matrix comprised of $q \times q$ blocks $\tilde{\mathbf{W}}_{ij}$ of dimension $p \times p$. Here $\mathbf{K}_{pq}$ is a vec-permutation matrix (Harville, 2008), namely

$$\mathbf{K}_{pq} = \sum_{i=1}^{p} \sum_{j=1}^{q} \mathbf{T}_{ij} \otimes \mathbf{T}'_{ij}$$

where $\mathbf{T}_{ij}$ is a $p \times q$ matrix whose (i, j) element is 1 and whose remaining elements are 0.

Proof of Theorem 3. First note that minimising equation (7.2) is equivalent to minimising the following function $l(\mathbf{U}, \mathbf{V})$ in terms of $\mathbf{U}$ and $\mathbf{V}$:

$$l(\mathbf{U}, \mathbf{V}) \quad = \quad \mathrm{tr}\{\mathbf{W}(\mathbf{V}^{-1} \otimes \mathbf{U}^{-1})\} - \log\det(\mathbf{V}^{-1} \otimes \mathbf{U}^{-1}). \tag{7.5}$$

This is, in fact, the log likelihood function of $(\mathbf{U}, \mathbf{V})$ given the sufficient statistics $\mathbf{W}$, up to a negative multiplicative constant. Thus the existence and the uniqueness of minimiser of l is equivalent to the existence and uniqueness of the maximum likelihood estimator of $(\mathbf{U}, \mathbf{V})$. Since a full rank matrix $\mathbf{W}$ can be regarded as a sample covariance matrix of data of size at least $n = pq$, the existence and the uniqueness up to a multiplicative constant follow directly from the necessary and sufficient condition given by Dutilleul (1999).

Write $\mathbf{\Lambda} \otimes \mathbf{U}$ as $(\lambda_{ij}\mathbf{U})$, $\mathbf{W}$ as $(\mathbf{W}_{ij})$, and apply matrix derivates. I then have

$$0 = \frac{\partial l}{\partial \mathbf{\Omega}} = \sum_{i,j} \lambda_{ij} \mathbf{W}_{ij} - p\mathbf{U}$$

which imples equation (7.3). Equation (7.4) can be obtained in a similar manner. If $v_{11} = 1$. the uniqueness of solution of equation (7.3) and (7.4) is implied by theorem 3.1 in Srivastava et al. (2008). □

Equations (7.3) and (7.4) have no analytic solutions but can be solved iteratively by using the following "flip-flop" algorithm where ϵ denotes an infinitesimal positive

quantitiy, $||\cdot||_2$ is the Euclidean norm and $\hat{\mathbf{V}}_0$ is an initial solution for $\mathbf{V}$. At each step, the current $\hat{\mathbf{V}}_c$ is used to compute the current $\hat{\mathbf{U}}_c$, which in turn is used to compute the next iterate of $\hat{\mathbf{V}}_n$ and then $\hat{\mathbf{U}}_n$ until $||\hat{\mathbf{V}}_n \otimes \hat{\mathbf{U}}_n - \hat{\mathbf{V}}_c \otimes \hat{\mathbf{U}}_c||_2 < \epsilon$. The Kullback-Leibler divergence in equation (7.2) decreases at each iteration, and the uniqueness of the solution of equations (7.3) and (7.4) guarantee that the numerical solution converge to a globe minimiser. Another question related to the algorithm is whether $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ are always positive definite in each iteration. The following theorem ensures the "flip-flop" algorithm will generate valid $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ at each iteration as long as the initial $\hat{\mathbf{V}}_0$ is positive definite.

Theorem 4. *If* $\mathbf{W}$ *and* $\mathbf{V}$ *are both positive definite matrices, then* $\mathbf{U}$ *obtained from equation (7.3) is also positive definite. The same is true for* $\mathbf{V}$ *in equation (7.4).*

Proof of Theorem 2. I only have to show that for any q-vector $\mathbf{x}$,

$$\mathbf{x}'(\frac{1}{p}\sum_{1\leq i,j\leq p}\lambda_{ij}\mathbf{W}_{i,j})\mathbf{x} > 0.$$

Notice that $\mathbf{x}'(\sum_{1\leq i,j\leq p}\lambda_{ij}\mathbf{W}_{i,j})\mathbf{x} = \mathrm{tr}(\mathbf{\Lambda Q}) > 0$ if $\mathbf{\Lambda}$ and $\mathbf{Q}$ are both positive definite where $\mathbf{Q} = (q_{ij})$ with $q_{ij} = \mathbf{x}'\mathbf{W}_{ij}\mathbf{x}$ is a symmetric matrix of dimension $p \times p$. $\mathbf{Q}$ is positive definite since for any p-vector y, $\mathbf{y}'\mathbf{Q}\mathbf{y} = \mathbf{y}'(1_q \otimes \mathbf{x}')\mathbf{W}(1_q \otimes \mathbf{x})\mathbf{y} > 0$. This accomplishs the proof. □

Now, any sample $\mathbf{W} \sim \pi$ on M_{pq}^+ can be projected to $\mathbf{V} \otimes \mathbf{U} \sim \pi_0$ on $M_p^+ \otimes M_q^+$ using the iteration algorithm.

Under the above compatible priors constructed by projection, B_{01} does not have an analytical expression, so requires numerical evaluation. One possible approach to compute B_{01} is to sample from the prior. In particular, if $\mathbf{W}_1, \cdots, \mathbf{W}_N$ is a sample from the prior π, then $N^{-1}\sum_{i=1}^N f\{\mathbf{Y} \mid h(\mathbf{W}_i)\}/N^{-1}\sum_{i=1}^N f(\mathbf{Y} \mid \mathbf{W}_i)$ estimates B_{01} consistently. But this estimate is usually quite poor, because the prior will not sample

intensively from the region where f is nonnegligible. This can be potentially improved by importance sampling. I can sample from a distribution of $\mathbf{W}$ with density $q(\mathbf{W})$, and then estimate the integral, for example, the numerator by: $N^{-1}\sum_{i=1}^{N} f\{\mathbf{Y} \mid h(\mathbf{W}_i)\}\pi(\mathbf{W}_i)/q(\mathbf{W}_i)$.

When a conjugate prior for unrestricted $\mathbf{W}$, such as an inverse Wishart prior, is used as $\pi(\mathbf{W})$, the denominator can be calculated in closed form; only the numerator requires approximation. One possible choice of importance distribution for the numerator would be the unrestricted posterior distribution $\pi(\mathbf{W} \mid \mathbf{Y})$, under which the Bayes factor can be simplified further as

$$B_{01} = \int_{M_{pq}^{+}} \frac{f\{\mathbf{Y} \mid h(\mathbf{W})\}}{f(\mathbf{Y} \mid \mathbf{W})} \pi(\mathbf{W} \mid \mathbf{Y}) \mathrm{d}\mathbf{W}.$$

The above Bayes factor for separability can be evaluated using Monte Carlo integration. I have experimented with examples that suggest potential usage of this Bayes factor and Monte Carlo approximations to test separability.

Extensions to graphical models are now trivial. For example, if $\mathbf{W}$ is constrained by a graph $G_{\mathbf{W}}$, the prior $\pi_1(\mathbf{W})$ can be adapted to the hyper-inverse Wishart prior corresponding to the graphical model restrictions. Moreover, if $\mathbf{U}$ and/or $\mathbf{V}$ are assumed to be constrained by graphs $G_{\mathbf{U}}$ and/or $G_{\mathbf{V}}$, the projection in Theorem 3 can be modified by using the theory of MLE of covariance matrices constrained by given graphs; then the corresponding prior distributions on $\mathbf{U}$ and $\mathbf{V}$ are compatible for Bayesian hypothesis testings against a non-separable $\mathbf{W}$.

7.2 Future work on dynamic seemingly unrelated regression models

The sparse seemly unrelated regression modelling approach described in Chapters 5 and 6 provides methodology for jointly modelling many regression coefficients and large-scale residual covariance matrices. The general goal of this approach is to help

build robust models and improve model interpretation in higher-dimensional regressions for multivariate responses. Sparsity - in terms of lower dimensional relationships underlying higher-dimensional patterns of associations - is key to achieving this goal. SSUR induces sparsity by selecting a subset of variables. I have demonstrated some of the utility of SSUR in applications of finance and econometrics, where identifying a subset of variables is practically useful.

Beyond variable selection, there are often situations that predicting future values of response variables, or understanding associations among predictors and/or response variables is of importance. It is then appealing to consider other types of lower dimensional structures that can help generate reliable inferences and predictions.

One potentially useful technique for imposing structures on seemingly unrelated regressions is the reduced rank model (Anderson, 1951; Reinsel & Velu, 1998). This model was originally developed for the multivariate regression models - a special type of seemingly unrelated regression models that assumes all response variables have the same predictors. Consider the multivariate linear regression models:

$$\mathbf{Y}_t' = \mathbf{F}_t'\boldsymbol{\Theta} + \boldsymbol{\nu}_t', \qquad \boldsymbol{\nu}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \tag{7.6}$$

for $t = 1, 2, \ldots,$ where

(i) $\mathbf{Y}_t = (Y_{ti})$ is the $p \times 1$ vector of response variables,

(ii) $\mathbf{F}_t$ is the known $n \times 1$ regressor vector,

(iii) $\boldsymbol{\Theta} = (\boldsymbol{\theta}_i)$ is the $n \times p$ matrix of regression coefficient matrix,

(iv) $\boldsymbol{\nu}_t = (\boldsymbol{\nu}_{ti})$ is the $p \times 1$ vector of random errors following multivariate normal distribution with mean vector $\mathbf{0}$, and covariance matrix $\boldsymbol{\Sigma}$.

In terms of scalar elements, I have p univariate models with the identical individual n-vector predictors, namely

$$\text{Observation:} \quad Y_{ti} = \mathbf{F}_t'\boldsymbol{\theta}_i + \boldsymbol{\nu}_{ti}, \qquad \boldsymbol{\nu}_{ti} \sim N(0, \sigma_{ii}^2) \tag{7.7}$$

for each i, t.

The reduced rank models approach the question of dimension reduction and structural modelling through the assumption of lower rank of matrix $\boldsymbol{\Theta}$ in model equation (7.6). More formally, assume that

$$\text{rank}(\boldsymbol{\Theta}) = r \leq \min(p, n).$$

This is equivalent to the parameter specification that $\boldsymbol{\Theta}$ can be expressed as

$$\boldsymbol{\Theta} = \boldsymbol{\Psi}\boldsymbol{\Phi}$$

where $\boldsymbol{\Psi}$ is of dimension $n \times r$ and $\boldsymbol{\Phi}$ is of dimension $r \times p$. The model of equation (7.6) can then be written as

$$\mathbf{Y}_t' = \mathbf{F}_t'\boldsymbol{\Psi}\boldsymbol{\Phi} + \boldsymbol{\nu}_t', \qquad \boldsymbol{\nu}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \tag{7.8}$$

where $\mathbf{F}_t'\boldsymbol{\Psi}$ is of reduced dimension with r components. Reduced rank regression models aim to use the r linear combinations of the predictor variables $\mathbf{F}_t$ to explain the variation in the response variables $\mathbf{Y}_t$. The practical implication is that there may not be a need for all n linear combinations or equivalently for all n predictors.

There is a rich literature on the inference in the reduced rank regression models. Some of the frequentist studies (Reinsel & Velu, 1998; Camba-Mendez et al., 2003) involve development of asymptotic sampling theory. MCMC based Bayesian analysis, and aspects of identification, prior specification, and model uncertainty with respect to the choice of rank r, appear in Geweke (1996a). Despite these advances, and the pressing need to develop methodology for higher-dimensional problems, these

works have generally focused on small problems - treating only a small number of response variables. One notable exception is Carriero et al. (2010), where a large set of response and predictive variables is considered. This paper provided a number of references to a large literature of frequentist and Bayesian analysis of multivariate regression models in the context of forecasting large datasets. It also compared a number of these methods for forecasting large datasets, and found that using shrinkage and rank reduction in combination rather than separately improve substantially the forecast accuracy. In the current methodology, latent association and parameter shrinkage are introduced through a two-stage technique for model fitting, no methodology exists that can formally reduce the rank and shrink parameters in a simultaneous manner. It is then appealing to consider a sparse reduced rank regression model that can potentially improve forecasting performance based on the synthesis of reduced rank and shrinkage techniques.

Another motivation for sparse reduced rank analysis comes from the desirable and inherent structure interpretation available in macroeconomic time series analysis. In such context, the predictors and the responses are both a large set of macroeconomic variables. I can image that combinations of predictors reflect individual aspects of economy: Financial market, real economy, and prices, etc. A given response macroeconomic variable may be only predicted by a few aspects of economy rather than predicted by all aspects of economy; so $\mathbf{\Phi}$ will have many zeros. Similarly, each economic aspect may involve a few economic variables but not all variables; so $\mathbf{\Psi}$ will have many zeros. This also motivates the integration of reduced rank modelling and sparse shrinkage modelling.

To improve both forecasting and interpretation, I want to propose the sparse reduced rank modeling that encourages sparse $\mathbf{\Phi}$ and $\mathbf{\Psi}$. A Bayesian approach to defining sparse reduced rank regression models can be developed in the spirit of sparse latent factor analysis (West, 2003; Carvalho et al., 2008). The general idea

is that I can use priors on the elements of $\mathbf{\Phi}$ and $\mathbf{\Psi}$ that induce zeros with high probability. Specifically, scalar elements $\Phi_{s,i}$ and $\Psi_{j,t}$ of $\mathbf{\Phi}_{s,i}$ and $\mathbf{\Psi}_{j,t}$, respectively, may be zero or take some non-zero value, for example, having the variable selection priors

$$\mathbf{\Phi}_{s,i} \sim \pi_i \delta_0(\mathbf{\Phi}_{s,i}) + (1 - \pi_i) N(\mathbf{\Phi}_{s,i} \mid 0, \sigma^2_{s,i})$$

where $\delta_0(\cdot)$ is the unit point mass at zero. Computationally, I expect that this general idea of sparse reduced rank regression models can be efficiently implemented by computational techniques such as MCMC algorithms. The resulting fully Bayesian analysis on large data sets can potentially offer a better prediction and easier structure interpretation than non-sparse reduced rank regression models.

Follow-on research to formally identify, or estimate, the rank r is of key interest. One possible approach for Bayesian model assessment of reduced rank regression involves the computation of posterior probability on the rank based on using MCMC methods for separate models differing only in the rank. Such a method requires the computation of marginal data densities as illustrated in Section 3.1 and 5.3.2. Various of methods are available for marginal data density computation. Lopes & West (2004) provided a wide ranging review of some methods in the context of latent factor analysis. Lopes & West (2004) also proposed a reversible jump Markov chain Monte Carlo algorithm to allow for uncertainty in the number of factors. These methods can be potentially used for rapid estimation of the number of the ranks, but may be sensitive to subjectively chosen priors. Another general strategy that seems very promising is to utilise a nonparametric Bayesian tool to automatically choose the rank and sparse pattern. This is related to the nonparametric infinite factor models of Bhattacharya & Dunson (2009), and approaches that put nonparametric process priors on latent factor model parameters.

Appendix A

Software manual for model implementation

In this appendix, I describe the user manual of computer codes that implement models and algorithms proposed throughout this thesis. Particularly, it contains the MCMC algorithms of Chapter 2 and 3 to fit matrix normal graphical model for given graphs and to explore graphical model uncertainty, the sequential graphical model search algorithm of Chapter 4, and the MCMC algorithms described in Chapter 5 and 6 to conduct sparse seemingly unrelated regression modelling. This software is freely available at my webpage `http://stat.duke.edu/~hw27`.

A.1 MCMC for matrix-variate graphical models

The directory "MatrixNormG" contains Matlab code and routines for the MCMC computations and model search in matrix-variate graphical models. Some of the data sets used in this thesis and the codes used to generate the results are described here.

Data File

The simulated example data in Chapter 2 and 3 is provided by the file "Eg1Simu.mat".

Using Code

Start with "DEMO.m" that illustrates the implementation of models using the simulated examples. Some of the key Matlab functions are given as follows:

Hyper-inverse Wishers sampler:

```
function [Omega,Sigma] = HIWsim(G,bG,DG,M)
%HIWSIM
% Samples the HIW_G(bG,DG) distribution on a graph G on p nodes
%Reference:  Carvalho,Massam and West (2007), Biometrika
p=size(DG,1);
Sigma = zeros(p,p,M); Omega=Sigma;     % arrays to save sampvar matrices
cliques = G{1}; separators = G{2};
numberofcliques = length(cliques);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Creat some working arrays that are computed only once
C1=inv(DG(cliques(1).ID,cliques(1).ID)/bG);
c1=cliques(1).ID; UN = c1';
for i=2:numberofcliques
    sid = separators(i).ID;   DSi{i}=inv(DG(sid,sid));
    cid = cliques(i).ID;      dif = setdiff(cid,UN) ;
    UN = union(cid',UN);
    sizedif = size(dif,2);
    DRS{i} = DG(dif,dif)-DG(dif,sid)*DSi{i}*DG(sid,dif);
    DRS{i}=(DRS{i}+DRS{i}')/2;
    mU{i}  = DG(dif,sid)*DSi{i};
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Now, MC Sampling
for j = 1:M
   UN = c1';
    Sigmaj=zeros(p,p);

    Sigmaj(c1,c1)=inv(wishart_InvA_rnd(bG+cliques(1).dim-1,...
        DG(cliques(1).ID,cliques(1).ID),1));
    % sample variance mx on first component
    for i=2:numberofcliques
% visit components and separators in turn
       dif = setdiff(cliques(i).ID,UN); UN = union(cliques(i).ID',UN);
       sizedif = size(dif,2); sid = separators(i).ID;
      SigRS = inv(wishart_InvA_rnd(bG+cliques(i).dim-1,DRS{i},1));
       Ui     = rMNorm(reshape(mU{i}',1,[]),kron(SigRS,DSi{i}),1);
       Sigmaj(dif,sid) = reshape(Ui,[],sizedif)'*Sigmaj(sid,sid);
       Sigmaj(sid,dif) = Sigmaj(dif,sid)';
       Sigmaj(dif,dif) = SigRS + ...
           Sigmaj(dif,sid)*inv(Sigmaj(sid,sid))*Sigmaj(sid,dif);
    end
```

```
    % Next, completion operation for sampled variance matrix
    H = c1;
    for i = 2:numberofcliques
        dif = setdiff(cliques(i).ID,H); sid = separators(i).ID;
        h = setdiff(H,sid);
        Sigmaj(dif,h) = Sigmaj(dif,sid)*inv(Sigmaj(sid,sid))...
            *Sigmaj(sid,h);
        Sigmaj(h,dif) = Sigmaj(dif,h)';
        H=union(H,cliques(i).ID);
    end
    Sigma(:,:,j)=Sigmaj;
    % Next, computing the corresponding sampled precision matrix
    Caux = zeros(p,p,numberofcliques); Saux = Caux;
    cid = cliques(1).ID; Caux(cid,cid,1) = inv(Sigmaj(cid,cid));
    for i = 2:numberofcliques
       cid = cliques(i).ID;       Caux(cid,cid,i) = inv(Sigmaj(cid,cid));
       sid = separators(i).ID;    Saux(sid,sid,i) = inv(Sigmaj(sid,sid));
    end
    Saux(:,:,1)=[]; % since we have separators indexed 2 up ...
    Omega(:,:,j) = sum(Caux,3) - sum(Saux,3);
end
% End of sampling
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Function for sampling inverse Wishart distribution conditional on the first element to be 1:

```
function f=iwishart_InvA_rnd_1(df,D)
%%Draw a sample f from conditional inv-wishart
%distribution IWishart(df) given f_11 =1


[p p]=size(D);
f=zeros(p);


D_21=D(2:end,1);
D_11=D(1,1);
D_22=D(2:end,2:end);
D_2dot1=D_22-D_21*inv(D_11)*D_21';
f_11=1;


f_2dot1=inv(wishart_InvA_rnd(df+p-1,D_2dot1,1) );

U=rMNorm(D_21*inv(D_11),inv(D_11)*f_2dot1,1);


f(1,1)=1;
f(2:end,1)=U*f(1,1);
f(1,2:end)=f(2:end,1)';
f(2:end,2:end)=f_2dot1+f(2:end,1)*inv(f(1,1))*f(1,2:end);
```

Function for sampling hyper-inverse Wishart distribution conditional on the first element to be 1:

```
function [Omega,Sigma] = HIWsim_con(G,bG,DG,M)

p=size(DG,1);
Sigma = zeros(p,p,M); Omega=Sigma;     % arrays to save sampvar matrices
cliques = G{1}; separators = G{2};
numberofcliques = length(cliques);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Creat some working arrays that are computed only once
C1=inv(DG(cliques(1).ID,cliques(1).ID)/bG);
c1=cliques(1).ID; UN = c1';
for i=2:numberofcliques
    sid = separators(i).ID;   DSi{i}=inv(DG(sid,sid));
    cid = cliques(i).ID;      dif = setdiff(cid,UN) ;
    UN = union(cid',UN);
    sizedif = size(dif,2);
    DRS{i} = DG(dif,dif)-DG(dif,sid)*DSi{i}*DG(sid,dif);
    DRS{i}=(DRS{i}+DRS{i}')/2;
    mU{i}  = DG(dif,sid)*DSi{i};
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Now, MC Sampling
for j = 1:M
   UN = c1';
   Sigmaj=zeros(p,p);
   Sigmaj(c1,c1)=iwishart_InvA_rnd_1(bG,DG(cliques(1).ID,cliques(1).ID));

    % visit components and separators in turn
   for i=2:numberofcliques
      dif = setdiff(cliques(i).ID,UN); UN = union(cliques(i).ID',UN);
      sizedif = size(dif,2); sid = separators(i).ID;
      SigRS = inv(wishart_InvA_rnd(bG+cliques(i).dim-1,DRS{i},1));

      Ui     = rMNorm(reshape(mU{i}',1,[]),kron(SigRS,DSi{i}),1);
      Sigmaj(dif,sid) = reshape(Ui,[],sizedif)'*Sigmaj(sid,sid);
      Sigmaj(sid,dif) = Sigmaj(dif,sid)';
      Sigmaj(dif,dif) = SigRS + Sigmaj(dif,sid)*inv(Sigmaj(sid,sid))...
          *Sigmaj(sid,dif);
   end
   % Next, completion operation for sampled variance matrix
   H = c1;
   for i = 2:numberofcliques
       dif = setdiff(cliques(i).ID,H); sid = separators(i).ID;
       h = setdiff(H,sid);
       Sigmaj(dif,h) = Sigmaj(dif,sid)*inv(Sigmaj(sid,sid))...
           *Sigmaj(sid,h);
       Sigmaj(h,dif) = Sigmaj(dif,h)';
```

```
        H=union(H,cliques(i).ID);
    end
    Sigma(:,:,j)=Sigmaj;
    % Next, computing the corresponding sampled precision matrix
    Caux = zeros(p,p,numberofcliques); Saux = Caux;
    cid = cliques(1).ID; Caux(cid,cid,1) = inv(Sigmaj(cid,cid));
    for i = 2:numberofcliques
       cid = cliques(i).ID;       Caux(cid,cid,i) = inv(Sigmaj(cid,cid));
       sid = separators(i).ID;    Saux(sid,sid,i) = inv(Sigmaj(sid,sid));
    end
    Saux(:,:,1)=[]; % since we have separators indexed 2 up ...
    Omega(:,:,j) = sum(Caux,3) - sum(Saux,3);
end
% End of sampling
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Gibbs sampler for matrix-variate graphical models

```
function [W_afterburn,V_afterburn,K_w_afterburn,...
    K_v_afterburn,loc_w_afterburn,loc_v_afterburn,v11,df_v11]...
    =cov_w_v_gibbs_hiwboth_star1(Y,b,B,d,D,GV,GW,adj,burnin,nmc)


[n,p,m]=size(Y);

%%% posterior
df_w=b+p*m;   df_v=d+n*m;   %D.F. of covariance W and V
df_kw=df_w+n-1;df_kv=df_v+p-1;

W_gibbs=zeros(n,n,nmc+burnin); % Set up variables to store samples
V_gibbs=zeros(p,p,nmc+burnin);


V_afterburn=zeros(p,p,nmc);
W_afterburn=zeros(n,n,nmc);


K_w_gibbs=zeros(n,n,nmc+burnin); % Set up variables to store samples
K_v_gibbs=zeros(p,p,nmc+burnin);

    loc_w_gibbs=zeros(n,n,burnin+nmc);
    loc_v_gibbs=zeros(p,p,burnin+nmc);

    loc_kw_gibbs=zeros(n,n,burnin+nmc);
    loc_kv_gibbs=zeros(p,p,burnin+nmc);

    v11=ones(1,nmc+burnin);

cliques = GV{1}; separators = GV{2};
numberofcliques = length(cliques);

jacob=(sum(sum(adj))+p)/2;

a_v11=cliques(1).dim*(d+2*cliques(1).dim)/2 ;
for i=2:numberofcliques
  a_v11=a_v11+cliques(i).dim*(d+2*cliques(i).dim)/2-separators(i).dim*...
      (d+2*separators(i).dim)/2;
end

a_v11 = a_v11-jacob;
df_v11=2*a_v11;

    V_gibbs(:,:,1)=eye(p); % Initializing V
     for i=2:burnin+nmc

      loc_w_gibbs(:,:,i)=sum_prod(Y,inv(V_gibbs(:,:,i-1)))+B;
      loc_kw_gibbs(:,:,i)=inv(loc_w_gibbs(:,:,i));

```

```
    % Draw a sample W from the conditional posterior
    [K_w_gibbs(:,:,i),W_gibbs(:,:,i)] = ...
        HIWsim(GW,df_w,loc_w_gibbs(:,:,i),1);


    v11(i)=1/gamrnd(a_v11,2/trace(D*inv(V_gibbs(:,:,i-1))));

    % Location Matrix of V given W
    loc_v_gibbs(:,:,i)=sum_prod_t(Y,K_w_gibbs(:,:,i))+D/v11(i);
    loc_kv_gibbs(:,:,i)=inv(loc_v_gibbs(:,:,i));
    %Draw a sample V from conditional posterior  given V_11 =1
    [K_v_gibbs(:,:,i),V_gibbs(:,:,i)] = ...
        HIWsim_con(GV,df_v,loc_v_gibbs(:,:,i),1);
    end

V_afterburn=V_gibbs(:,:,burnin+1:end);
W_afterburn=W_gibbs(:,:,burnin+1:end);

K_v_afterburn=K_v_gibbs(:,:,burnin+1:end);
K_w_afterburn=K_w_gibbs(:,:,burnin+1:end);

    loc_w_afterburn=loc_w_gibbs(:,:,burnin+1:end);
    loc_v_afterburn=loc_v_gibbs(:,:,burnin+1:end);
%
```

A.2 Sequential stochastic search for dynamic graphical models

Fixed graphs

The directory "SSS_type1" contains a C++ program that uses the sequential stochastic search of Section 4.4.2 to sequentially sample decomposable graphs according to their predicted probability described by equation. 4.8. This manual describes the compilation of the program, the input parameters to be set by the user in several input files, the output parameters that are saved in files specified by users, and other parameters interested users can modify by changing and recompiling the source codes.

COMPILATION: Type "make".

INPUT: The following parameters are specified by the user and read in via the standard input (eg $./main < infile$). See the "infile" given for an example.

(i) Residual file: Character string. The residual file stores the sequence of standardised residuals, $\mathbf{e}_t/\sqrt{q}_t$, from the output of DLM models.

(ii) $\mathbf{q}_t$ file: Character string. The $\mathbf{q}_t$ file stores the sequence of q_t's from the output of inferences on DLM models.

(iii) Initial adjacency matrix file: Character string. The $p \times p$ adjacency matrix is specified by a sequence of 0's and 1's of length $p^2 \times 1$. These are the elements of the adjacency matrix, listed row by row.

(iv) Prior hyper-parameter $\mathbf{S}_0$ file: Character string. The $p \times p$ matrix, $\mathbf{S}_0$, is specified by a column of length $p^2 \times 1$. These are the elements of the $\mathbf{S}_0$ matrix, listed row by row.

(v) Number of stocks, i.e. the value of p.

(vi) Number of time series observations.

(vii) Value of the prior hyper parameter b_0.

(viii) Number of top graphs used to conduct Bayesian model averaging.

(ix) Number of maximum shotgun stochastic search iterations at each time point.

(x) Value of the threshold used to stop shotgun stochastic search at each time point.

(xi) Value of the starting time point to begin the sequential learning of graphical models.

(xii) Value of the discount factor δ.

(xiii) Value of the annealing parameter c.

OUTPUT: The following output files will be created to record the results from the sequential analysis:

(i) Log predictive density file: Character string. The log predictive density is sequentially computed based on equation 4.12

(ii) Predicted residual covariance matrix file: Character string. The predicted residual covariance matrix is the quantity of $E(\mathbf{\Sigma}_t \mid I_t)$ in Theorem 2.

(iii) Predicted edge inclusion probability file: Character string. The edge inclusion probability at each time point t is computed according to equation 4.9.

(iv) File for recording the number of shotgun stochastic search iterations for each time point: Character string.

(v) File for recording the maximum absolute difference between correlation matrices from the last two shotgun iterations.

The paths and names of these files are specified by user in the "infile" and read in via the standard input.

OTHER PARAMETERS: Users can change the hyper-parameters of the sparse encouraging prior over graphs by modifying values at line 76 of the "main.C" file.

Discounted graphs

The directory "SSS_type2" contains a C++ program that uses the sequential stochastic search of Section 4.4.2 to sequentially sample decomposable graphs according to their predicted probability described by equation 4.9. The manual is the same as those for the program in "SSS_type1".

A.3 Matlab codes for sparse seemingly unrelated regression models

The directory "SSURDemo" contains Matlab code and routines for the MCMC computations and model search in SSUR models. Some of the data sets used in this thesis and the codes used to generate the results are described as follows.

Data File

The example data is provided under the directory Data:

- "F-F_Research_Data_Factors.txt": This file contains data for extracting month information useful for VAR example;
- "Macro_MHsearch.mat": This file contains results of the analysis of the VAR example in the paper;
- "Macro_MHsearch.mat": This file contains results of the analysis of the VAR example in the paper;
- "Simu_p6": This file contains simulated data in the simulated example;
- "varex": this file contains macroeconomic data for the VAR example.

Using Code

Start with "main_macro.m" or "main simu.m" that illustrate the implementation of models using the first simulated example and the first real-world example:

- "MultiLinearSamplerMarglik.m": This function performs the Gibbs sampling and marginal likelihood approximation for fixed adjacency matrix specified by variable *adj* and subset of variables specified by variable *z1*;
- "SUR_SSVS.m": This function indirectly sample the model space without computing marginal likelihood;

- "SSUR_MH": This function searches for SSUR model using the two approximations for marginal data densities.

Bibliography

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics* **22**, 327–351.

Ando, T. & Zellner, A. (2010). Hierarchical Bayesian analysis of the seemingly unrelated regression and simultaneous equations models using a combination of direct Monte Carlo and importance sampling techniques. *Bayesian Analysis* **5**, 847–878.

Besag, J. (1989). A Candidate's formula: A curious result in Bayesian prediction. *Biometrika* **76**, 183–183.

Bhattacharya, A. & Dunson, D. B. (2009). Sparse Bayesian infinite factor models. Tech. rep., Department of Statistical Science, Duke University, Discussion Papers.

Box, G. E. P. & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley.

Brown, P. J., Vannucci, M. & Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **60**, 627–641.

Camba-Mendez, G., Kapetanios, G., Smith, R. J. & Weale, M. R. (2003). Tests of rank in reduced rank regression models. *Journal of Business and Economic Statistics* **21**, 145–155.

Carriero, A., Kapetanios, G. & Marcellino, M. (2010). Forecasting large datasets with Bayesian reduced rank multivariate models. *Journal of Applied Econometrics (to appear)* .

Carvalho, C. M. (2006). *Structure and Sparsity in High-Dimensional Multivariate Analysis*. PhD in Statistical Science, Department of Statistical Science–Duke University.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q. & West, M. (2008). High-dimensional sparse factor modeling: Applications in gene

expression genomics. *Journal of the American Statistical Association* **103**, 1438–1456.

Carvalho, C. M., Massam, H. & West, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika* **94**, 647–659.

Carvalho, C. M. & Scott, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96**, 497–512.

Carvalho, C. M. & West, M. (2007a). Dynamic matrix-variate graphical models. *Bayesian Analysis* **2**, 69–98.

Carvalho, C. M. & West, M. (2007b). Dynamic matrix-variate graphical models - A synopsis. In *Bayesian Statistics VIII*, J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith & M. West, eds. Oxford University Press, pp. 585–590.

Chen, Z. & Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59**, 762–769.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313–1321.

Chib, S. & Greenberg, E. (1995). Hierarchical analysis of SUR models with extensions to correlated serial errors and time-varying parameter models. *Journal of Econometrics* **68**, 339 – 360.

Cripps, E., Carter, C. K. & Kohn, R. (2005). Variable selection and covariance selection in multivariate regression models. In *Handbook of Statistics 25 Bayesian Thinking: Modeling and Computation*, D. K. Dey & C. R. Rao, eds. Elsevier.

Daniels, M. J. & Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* **89**, 553–566.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68**, 265–274.

Dawid, A. P. & Lauritzen, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics* **21**, 1272–317.

Dawid, A. P. & Lauritzen, S. L. (2000). Compatible prior distributions. In *Bayesian Methods with Applications to Science, Policy, and Official Statistics Selected Papers from ISBA 2000:The Sixth World Meeting of the International Society for Bayesian Analysis*, E. I. George, ed. Eurostat.

Dellaportas, P. & Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86**, 615–633.

Dempster, A. (1972). Covariance selection. *Biometrics* **28**, 157–75.

Dobra, A., Jones, B., Hans, C., Nevins, J. & West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis* **90**, 196–212.

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation* **64**, 105–123.

Fama, E. F. (1981). Stock returns, real activity, inflation, and money. *The American Economic Review* **71**, 545–565.

Fama, E. F. & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* **33**, 3–56.

Finn, J. D. (1974). *A General Model for Multivariate Analysis*. New York: Holt, Rinehart and Winston.

Foschi, P., Belsley, D. A. & Kontoghiorghes, E. J. (2003). A comparative study of algorithms for solving seemingly unrelated regressions models. *Computational Statistics and Data Analysis* **44**, 3 – 35.

Foschi, P. & Kontoghiorghes, E. J. (2002). Seemingly unrelated regression model with unequal size observations: Computational aspects. *Computational Statistics & Data Analysis* **41**, 211 – 229.

Galecki, A. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics - Theory and Methods* **23**, 3105 – 3119.

Gelfand, A. E. & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)* **56**, 501–514.

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association* **99**, 537–545.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **3**, 515–534.

George, E. I. & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

George, E. I. & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–379.

George, E. I., Sun, D. & Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics* **142**, 553 – 580.

Geske, R. & Roll, R. (1983). The fiscal and monetary linkage between stock returns and inflation. *The Journal of Finance* **38**, 1–33.

Geweke, J. (1996a). Bayesian reduced rank regression in econometrics. *Journal of Econometrics* **75**, 121 – 146.

Geweke, J. F. (1996b). Variable selection and model comparison in regression. In *Bayesian Statistics 5*, J. M. Bernado, J. O. Berger, A. P. Dawid & A. F. M. Smith, eds. Oxford Univeristy Press, pp. 609–620.

Giudici, P. (1996). Learning in graphical Gaussian models. In *Bayesian Statistics 5*, J. M. Bernado, J. O. Berger, A. P. Dawid & A. M. Smith, eds. Oxford Univeristy Press, pp. 621–628.

Giudici, P. & Castelo, R. (2003). Improving Markov chain Monte Carlo model search for data mining. *Machine Learning* **50**, 127–158.

Giudici, P. & Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86**, 785–801.

Griffiths, W. (2001). Bayesian inference in the seemingly unrelated regressions models. Department of economics - working papers series, The University of Melbourne.

Gupta, A. K. & Nagar, D. K. (2000). *Matrix Variate Distributions*, vol. 104 of *Monographs and Surveys in Pure & Applied Mathematics*. London: Chapman & Hall.

Hammersley, J. M. & Clifford, P. E. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.

Hans, C. (2005). *Regression Model Search and Uncertainty with Many Predictors*. PhD in Statistical Science, Institude of Statistics and Decision Sciences–Duke University.

Hans, C., Dobra, A. & West, M. (2007). Shotgun stochastic search in regression with many predictors. *Journal of the American Statistical Association* **102**, 507–516.

Harville, D. A. (2008). *Matrix Algebra from a Statistician's Perspective*. New York: Springer-Verlag.

Hobert, J. P. & Marchev, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *Annals of Statistics* **36**, 532–554.

Holmes, C. C., Denison, D. G. T. & Mallick, B. K. (2002). Accounting for model uncertainty in seemingly unrelated regressions. *Journal of Computational and Graphical Statistics* **11**, 533–551.

Huizenga, H. M., de Munck, J. C. & Waldorp, L. J. Grasman, R. (2002). Spatiotemporal EEG/MEG source analysis based on a parametric noisecovariance model. *IEEE Transactions on Biomedical Engineering* **49**, 533–539.

Jones, B., Carvalho, C. M., Dobra, A., Hans, C., Carter, C. & West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science* **20**, 388–400.

Jordan, M., Ghahramani, Z., Jaakkola, T. & Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning* **37**, 183–233.

Kontoghiorghes, E. J. & Clarke, M. R. B. (1995). An alternative approach for the numerical solution of seemingly unrelated regression equations models. *Computational Statistics and Data Analysis* **19**, 369–377.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.

Lee, B.-S. (1992). Causal relations among stock returns, interest rates, real activity, and inflation. *The Journal of Finance* **47**, 1591–1603.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics* **47**, 13–37.

Liu, C., Rubin, D. B. & Wu, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* **85**, 755–770.

Liu, J. S. & Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association* **94**, 1264–1274.

Lopes, H. & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41–67.

Lu, N. & Zimmerman, D. L. (2005). The likelihood ratio test for a separable covariance matrix. *Statistics and Probability Letters* **73**, 449–457.

Madigan, D. & York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63**, 215–232.

Mardia, K. V. & Goodall, C. R. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. In *Multivariate Environmental Statistics*, G. P. Patil & C. R. Rao, eds. Elsevier, pp. 347–385.

Markowitz, H. (1959). *Portfolio Selection: Efficient Diversification of Investments*. New York, USA: John Wiley and Sons.

McCulloch, R., Polson, N. & Rossi, P. (2000). Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics* **99**, 173–193.

McCulloch, R. E. & Rossi, P. E. (1992). Bayes factors for nonlinear hypotheses and likelihood distributions. *Biometrika* **79**, 663–676.

Meng, X.-L. & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica* **6**, 831–860.

Min, C. & Zellner, A. (1993). Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *Journal of Econometrics* **56**, 89–118.

Mitchell, M. W., Genton, M. G. & Gumpertz, M. L. (2005). Testing for separability of space-time covariances. *Environmetrics* **16**, 819–831.

Mitchell, M. W., Genton, M. G. & Gumpertz, M. L. (2006). A likelihood ratio test for separability of covariances. *Journal of Multivariate Analysis* **97**, 1025–1043.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica* **34**, 768–783.

Naik, D. N. & Rao, S. S. (2001). Analysis of multivariate repeated measures data with a Kronecker product structured covariance matrix. *Journal of Applied Statistics* **29**, 91–105.

Pástor, L. & Stambaugh, R. F. (2002). Mutual fund performance and seemingly unrelated assets. *Journal of Financial Economics* **63**, 315 – 349.

Pole, A., West, M. & Harrison, P. (1994). *Applied Bayesian Forecasting and Time Series Analysis*. New York: Chapman-Hall.

Quintana, J. (1992). Optimal portfolios of forward currency contracts. In *Bayesian Statistics IV*, J. Berger, J. Bernardo, A. Dawid & A. Smith, eds. Oxford University Press, pp. 753–762.

Quintana, J., Lourdes, V., Aguilar, O. & Liu, J. (2003). Global gambling. In *Bayesian Statistics VII*, J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith & M. West, eds. Oxford University Press, pp. 349–368.

Quintana, J. M., Carvalho, C. M., Scott, J. & Costigliola, T. (2009). Futures markets, Bayesian forecasting and risk modeling. In *The Handbook of Applied Bayesian Analysis*, A. O'Hagan & M. West, eds. Oxford University Press, pp. 343–365.

Quintana, J. M. & West, M. (1987). Multivariate time series analysis: New techniques applied to international exchange rate data. *The Statistician* **36**, 275–281.

Raftery, A. E., Madigan, D. & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 1197–1208.

Rajaratnam, B., Massam, H. & Carvalho, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *Annals of Statistics* **36**, 2818–49.

Reinsel, G. C. & Velu, R. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*. New York: Springer.

Roy, V. & Hobert, J. P. (2007). Convergence rates and asymptotic standard errors for MCMC algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society, (Series B)* **69**, 607–623.

Scott, J. G. & Carvalho, C. M. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics* **17**, 790–808.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* **19**, 425–442.

Smith, M. & Kohn, R. (2000). Nonparametric seemingly unrelated regression. *Journal of Econometrics* **98**, 257 – 281.

Smith, M. & Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association* **97**, 1141–1153.

Srivastava, M. S., von Rosen, T. & von Rosen, D. (2008). Models with a Kronecker product covariance structure: Estimation and testing. *Mathematical Methods of Statistics* **17**, 357–370.

Srivastava, V. K. & Giles, D. E. A., eds. (1987). *Seemingly Unrelated Regression Equations Models*. New York, NY, USA: Marcel Dekker, Inc.

Theobald, D. L. & Wuttke, D. S. (2006). Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian procrustes problem. *Proceedings of the National Academy of Sciences* **103**, 18521–18527.

Wang, H., Reeson, C. & Carvalho, C. M. (2009). Dynamic Financial Index Models: Modeling conditional dependencies via graphs. Tech. rep., Department of Statistical Science, Duke University, Discussion Papers.

Wang, H. & West, M. (2009). Bayesian analysis of matrix normal graphical models. *Biometrika* **96**, 821–834.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics 7*, J. M. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith & M. West, eds. Oxford University Press.

West, M. & Harrison, P. (1997). *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag, 2nd ed.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester UK: John Wiley and Sons.

Wong, F., Carter, C. & Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika* **90**, 809–30.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* **57**, 348–368.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

Zellner, A. & Chetty, V. K. (1965). Prediction and decision problems in regression models from the Bayesian point of view. *Journal of the American Statistical Association* **60**, 608–616.

Biography

Hao Wang was born in Shaoxing, ZJ, China on November 18, 1983, and attended high school at Shaoxing No. 1 high school. After graduating from high school in 2002, he enrolled at Nankai University in Tianjin, China, where he studied applied mathematics before switching to statistics. He received his bachelor's degree in June of 2006. He then went to the Department of Statistical Science at Duke University as a graduate student. At Duke, he worked on Bayesian statistics, and proudly became a Bayesian. In 2008, he earned a M.S. in Statistical Science from Duke.