STRUCTURE AND SPARSITY IN HIGH-DIMENSIONAL MULTIVARIATE ANALYSIS

by

Carlos Marinho Carvalho

Institute of Statistics and Decision Sciences
Duke University

Date: ______________________________

Approved:

____
Dr. Mike West, Supervisor

____
Dr. Alan Gelfand

____
Dr. Merlise Clyde

____
Dr. David Dunson

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Institute of Statistics and Decision Sciences
in the Graduate School of
Duke University

February 2006

ABSTRACT

(Statistics)

STRUCTURE AND SPARSITY IN HIGH-DIMENSIONAL MULTIVARIATE ANALYSIS

by

Carlos Marinho Carvalho

Institute of Statistics and Decision Sciences
Duke University

Date: ____________________
Approved:

____
Dr. Mike West, Supervisor

____
Dr. Alan Gelfand

____
Dr. Merlise Clyde

____
Dr. David Dunson

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Institute of Statistics and Decision Sciences in the Graduate School of
Duke University

February 2006

Abstract

As scientific problems grow in terms of both expanding parameter dimension and sample sizes, structure and sparsity become central concepts in practical data analysis and inference. By allowing complex high-dimensional problems to be modeled through low-dimensional underlying relationships, sparsity helps to simplify estimation, reduce computational burden and facilitate interpretation of large scale datasets. This dissertation addresses the issue of sparsity modeling primarily in the context of Gaussian graphical models and sparse factor models.

Chapter 1 contextualizes the dissertation by introducing the way sparsity models are discussed throughout this work.

Chapter 2 introduces the basic theory of Gaussian graphical models and central elements of Bayesian analysis in this class of models.

Chapter 3 is concern with problem of model determination in graphical model space. Existing methods are tested in high-dimensional setups and a novel parallel stochastic search method is described. Both decomposable and non-decomposable graphs are considered. Examples of moderate (12-20) to large (150) size are considered, combining simple synthetic examples with data analysis from gene expression studies.

Chapter 4 develops a efficient method for direct simulation from the hyper-inverse Wishart prior/posterior on any defined graphical model. This new sampling method provides completion of the simulation toolbox for Bayesian exploration and analysis of Gaussian graphical models under HIW priors.

Chapter 5 extends conditional independence ideas from Gaussian graphical models to multivariate dynamic linear models. After presenting the development

of this new class of models the chapter focuses on applications of such models in large financial time series portfolio allocation problems.

Chapter 6 deals with sparse factor models where model search and fitting are addressed through stochastic simulation (MCMC) and a novel computational strategy involving a evolutionary search to address the issue of identifying variables for inclusion. This forms a first, Bayesian “projection pursuit” method relevant in high-dimensional factor and structure analysis. Examples are drawn from genomic studies where factor models aim to identify multi-dimensional biological patterns related to oncogenic pathways.

Finally, Chapter 7 summarizes the dissertation and discusses possible generalizations and future work.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to acknowledge my friends and colleagues at ISDS who have supported and helped me over the past few years. I would like to especially thank my advisor, Mike West, for his guidance, encouragement and constant support during the development of this dissertation and my years as a Ph.D. student. Throughout the years, Mike was not only a mentor but also a colleague and a friend. His excitement about and commitment to research set a strong example that I hope to emulate as I grow in academia.

I would like to thank Merlise Clyde for many, many hours of invaluable discussions on a variety of topics that have helped me move forward. I thank Beatrix Jones and Adrian Dobra for their important role as collaborators in my early years at ISDS and Quanli Wang for all his help with computational issues. I also thank Hedibert Lopes for all his guidance, before and after I arrived at Duke.

A special thanks goes to Chris Hans for being an inspiring colleague, a formidable golf partner, and above all, a great friend.

Thanks to my fabulous wife Jill for all her patience, love and enthusiasm about every little thing in life. Everything is better with you beside me.

Finally, I would like to thank my parents. It is almost impossible to describe in words the endless support and unconditional love that they have given me on every step of the way. To you, Carlos and Tania, I say muito obrigado and dedicate this work.

Chapter 1

Introduction

Constant technological advances have dramatically increased, in recent years, our capacity to collect and store enormous amount of data. From high-frequency finance and very large marketing datasets to high throughput genomic data we face applied problems of increasing dimension. The necessity to analyze, interpret and extract relevant information from these large datasets has pushed statistical sciences into a "revolution" where the design of higher-dimensional models and associated computational tools is a focal point of research. The impact of applied Bayesian analysis has been particularly notable as development of stochastic simulation methods enable the application of more complex and more realistic mathematical models. The use of stochastic computational methods for inference in large parametric spaces and model determination raises a number of challenges of both statistical and computational efficiency as well as basic feasibility.

The imposition of structure is an essential step in dealing with large problems, as it helps to simplify estimation, reduce computational burden and facilitate interpretation. In general, this concept is a way of imposing constraints that force objects of interest to lie in lower-dimensional spaces, in line with the scientific view of parsimony. Structure, however, comes with a price as the task of determining its form might represent a daunting computational effort. With these ideas in mind, this dissertation addresses methodological and computational aspects of structured high-dimensional multivariate problems with inferential procedures and model specification tools being developed. Throughout this work, structure comes in the form of sparsity in terms of low-dimensional relationships underlying high-dimensional patterns of association. More specifically, we focus on models for covariance matrices defined via parametric and conditional independence constraints with applications ranging from financial econometrics to functional genomics studies.

The initial chapters of this dissertation concentrate on Gaussian graphical models (Whittaker, 1990; Lauritzen, 1996) where nodes represent random variables and the set of edges define the global conditional independence structure of the distribution. Graphical structures provide computational efficiencies and visual aids by a decomposition of the sample space into subsets of variables, reducing the problem to a collection of typically small (local) models. Identifying interesting graphs under the implied posterior distribution presents us with a hard challenge as the number of possible models grows prohibitively large with dimension. To address this, Chapters 2 and 3 explore scalability and efficacy of existing stochastic computational tools for model determination in graphical model space, along with the development of a novel parallel stochastic search algorithm. Consistency of prior specification as well as sparsity priors over graphs are discussed, together with the presentation of a central element of Bayesian analysis in graphical models: the hyper inverse-Wishart distribution (HIW). Unlike many recent efforts in the literature, non-decomposable graphical models are considered in all search schemes presented which can be viewed as my main contribution in this research area. Chapter 4 completes the full simulation-based Bayesian analysis framework of Gaussian graphical models by developing an efficient method for direct simulation of the hyper inverse-Wishart distribution on any graph, where the use of the junction tree of a graph allow us to work sequentially at the prime component level, avoiding the manipulation of large matrices.

Inspired by results of a portfolio allocation example in Chapter 4, in Chapter 5, I extend conditional independence ideas to dynamic linear models (West and Harrison, 1997), defining a new class of multivariate DLMs where hyper inverse-Wishart distributions are used to model the cross-sectional covariance structure of a set of time series. High-dimensional sequential portfolio allocation examples display highly encouraging results indicating that proper handling of conditional independence among assets is fundamental in portfolio theory, raising new questions in that field and highlighting the utility of sparse modeling.

The second part of the dissertation, Chapter 6, deals with sparse factor models (West, 2003). Sparsity has a new meaning in this context as zeros in the factor loadings matrix promote constraints in the covariance matrix rather than its inverse, as in graphical models. One key motivating application context for these models is the identification of complex multi-dimensional genomic patterns related to deregulation of oncogenic pathways. Here, factor analysis aims to decompose the variation of large dimensional gene expression datasets as an attempt to improve our understanding of the genomics and genetics of these pathways. Once again, sparsity plays a fundamental role in helping reduce the parametric space and providing a formal model for pathway interpretation. This chapter describes the MCMC methodology for inference in sparse factor models including the adaptation of a novel hierarchical shrinkage prior (Lucas *et al.*, 2006) that act more aggressively than traditional priors in isolating signal from noise. My contribution to this research is a key methodological development concerning an evolutionary model determination process that sequentially expands the dimension of the sample space, enriching the analysis of existing factors. In our genomics studies, it is often the case that the exploration of a particular pathway starts from a list of know genes involved in a biological process and expanding the analysis by including genes showing association with the initial set of variables is a natural step. Critical issues in model specification such as identification constraints and choice of the number of factors are also a target of the evolutionary search. A comprehensive simulated example tests the performance of the evolutionary search whereas an example involving a very important hormonal oncogenic pathway (ER pathway) illustrates the methodology as an approach to explore, evaluate and define molecular phenotypes.

Throughout this dissertation theoretical and methodological aspects of large multivariate problems are discussed along with the development of innovative computational tools for model selection and inference. Extensions of the material presented here include more developments in model selection in graphical model space, better understating of the implications of conditional independence constraints in investment decisions and the further exploration of biological pathways via sparse factor models. In Chapter 7, I conclude this work with the discussion of some open questions and follow-up goals that are subject of my current research agenda.

Chapter 2

Gaussian Graphical Models

A graphical model is a probability model that characterizes the conditional independence structure of a set of random variables by a graph (Whittaker, 1990; Lauritzen, 1996). Graphs provide a way to decompose the sample space into subsets of variables (graph vertices) generating efficient ways to model conditional and marginal distributions locally so that complex problems can be handled through the combination of simple elements. In high-dimensional problems, graphs are a natural way to reduce the parameter space, a fundamental step in modeling situations where the number of variables exceeds the number of observations.

In the context of a multivariate normal distribution, conditional independence restrictions are simply expressed through zeros in the off-diagonal elements of the precision matrix (inverse of the covariance matrix, also known as the concentration matrix), establishing a parsimonious way to model covariance structures. This approach dates back to Dempster (1972) in the so-called covariance selection models and, after associations to graphical ideas made by Speed and Kiiveri (1986), the term Gaussian graphical models becomes standard. The introduction of the hyper-inverse Wishart (Dawid and Lauritzen, 1993) as a conjugate prior for structured covariance matrices aids the development of Bayesian approaches to covariance matrix estimation and graphical model selection.

This chapter focus on the description of central concepts of Gaussian graphical models. It starts with a brief presentation of graphs and conditional independence ideas, followed by elements of Bayesian analysis in this class of models.

2.1 Basic Graph Theory

2.1.1 Notation

A graph is a visual object defined by the pair (V, E) where V is the vertex set of p elements (variables) and E defines the edge-set. A graph can be represented by a picture where each vertex is a circle with arrows or lines displaying the edges in E. Edge $(i, j) \in E$ is called an *undirected* edge if (j, i) is also in E and is represented by a line connecting vertex i to vertex j. If $(i, j) \in E$ but its opposite (j, i) is not, (i, j) is called a *directed* edge. If a graph has only undirected edges it is called *undirected graph* whereas if all edges are directed the graph is said to be a *directed graph.*

Let $G = (V, E)$ be an undirected graph. Vertices a and b are said to be *neighbors* (or adjacent) in G if there is an edge $(a, b) \in E$. A graph (or subgraph) is *complete* if all of its vertices are connected by edges in E. A *clique* is a complete subgraph that is not contained within another complete subgraph. Subgraphs (A, B, C) form a decomposition of G if $V = A \cup B$, $C = A \cap B$ is complete and C *separates* A from B (any path from A to B goes to through C). The subgraph C is said to be a *separator.* A sequence of subgraphs that cannot be further decomposed are the *prime components* of a graph. A graph is said to be *decomposable* if every prime component is complete.

A graph G can be represented by a *perfect ordering* of its prime components and separators. A ordering of components $P_i \in \mathcal{P}$ and separators $S_i \in \mathcal{S}$, $(P_1, S_2, P_2, S_3, \ldots, P_k)$, is said to be perfect if for every $i = 2, 3, \ldots, k$ there exists a $j < i$ such that

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

where

$$H_{i-1} = \bigcup_{j=1}^{i-1} P_j.$$

Any connected graph G can be represented as a tree of its prime components – the *junction tree*. A tree with set of vertices equal to the the set of prime components of G is said to be a junction tree if for any two prime components P_i and P_j and any T on the unique path between P_i and P_j, $P_i \cap P_j \subset T$. A set of vertices shared by two adjacent nodes of the junction tree is complete and defines the *separator* of the two subgraphs induced by the nodes. This representation plays a critical role in computational aspects of graphs and most of the methodology developed and explored in this dissertation depend on it. An example of a decomposable graph and its junction tree is presented in Figure 2.1 and efficient algorithms for producing the junction tree for any given graph are presented in Appendix A and illustrated in a non-decomposable graph in Figure 2.2.

2.1.2 Conditional Independence and Markov Properties

A graph is a simple way to summarize a collection of marginal and conditional independences in a joint probability distribution over a collection of variables.

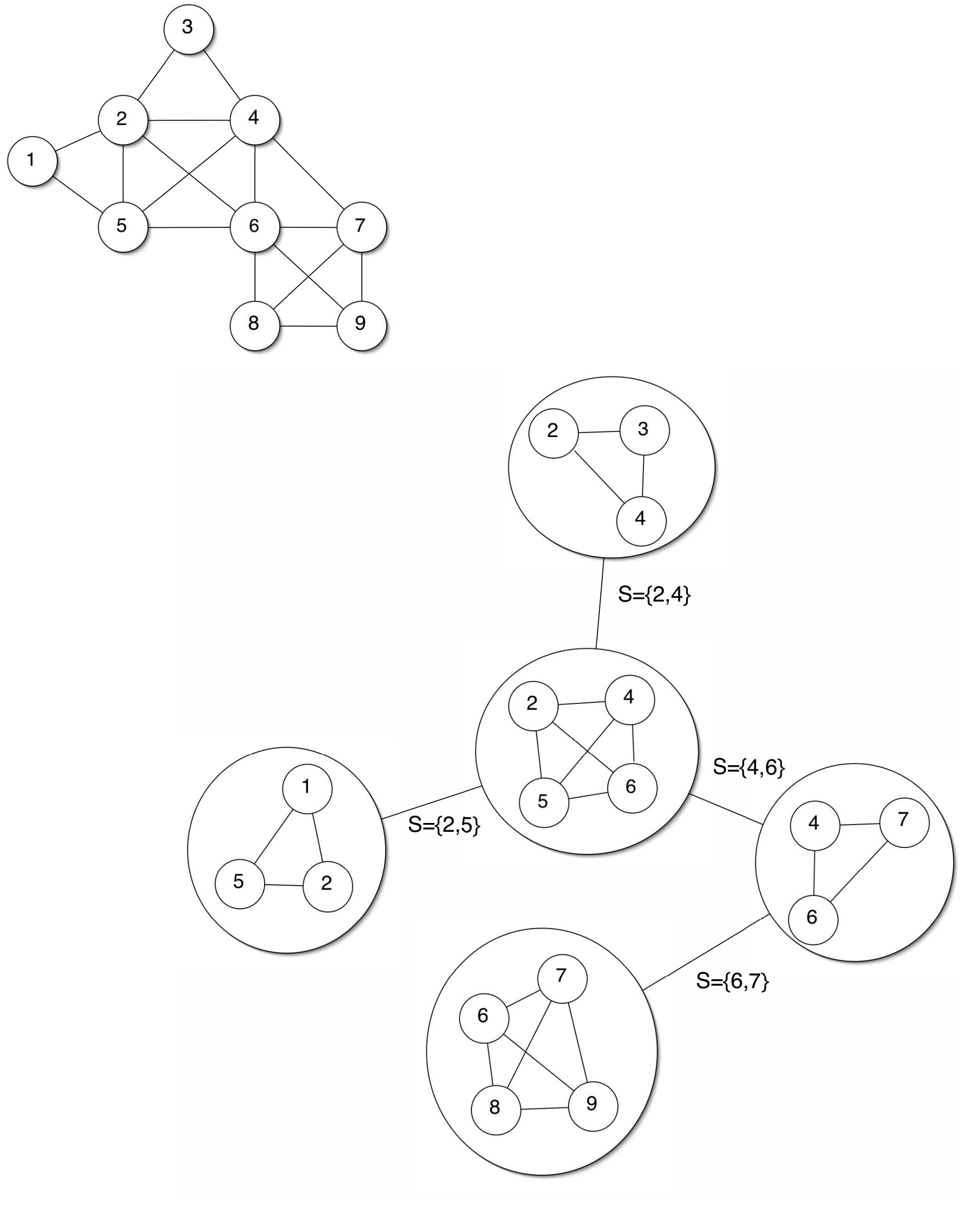


Figure 2.1: A decomposable graph (top frame) and its junction tree (bottom frame). Each node of the junction tree represents a clique while vertices shared by adjacent nodes of the tree define the separators. In this graph, $\{\{1, 2, 5\}, \{2, 4, 5, 6\}, \{2, 3, 4\}, \{4, 6, 7\}, \{6, 7, 8, 9\}\}$ is the set of cliques and $\{\{2, 5\}, \{2, 4\}, \{4, 6\}, \{6, 7\}\}$ is the separators set.

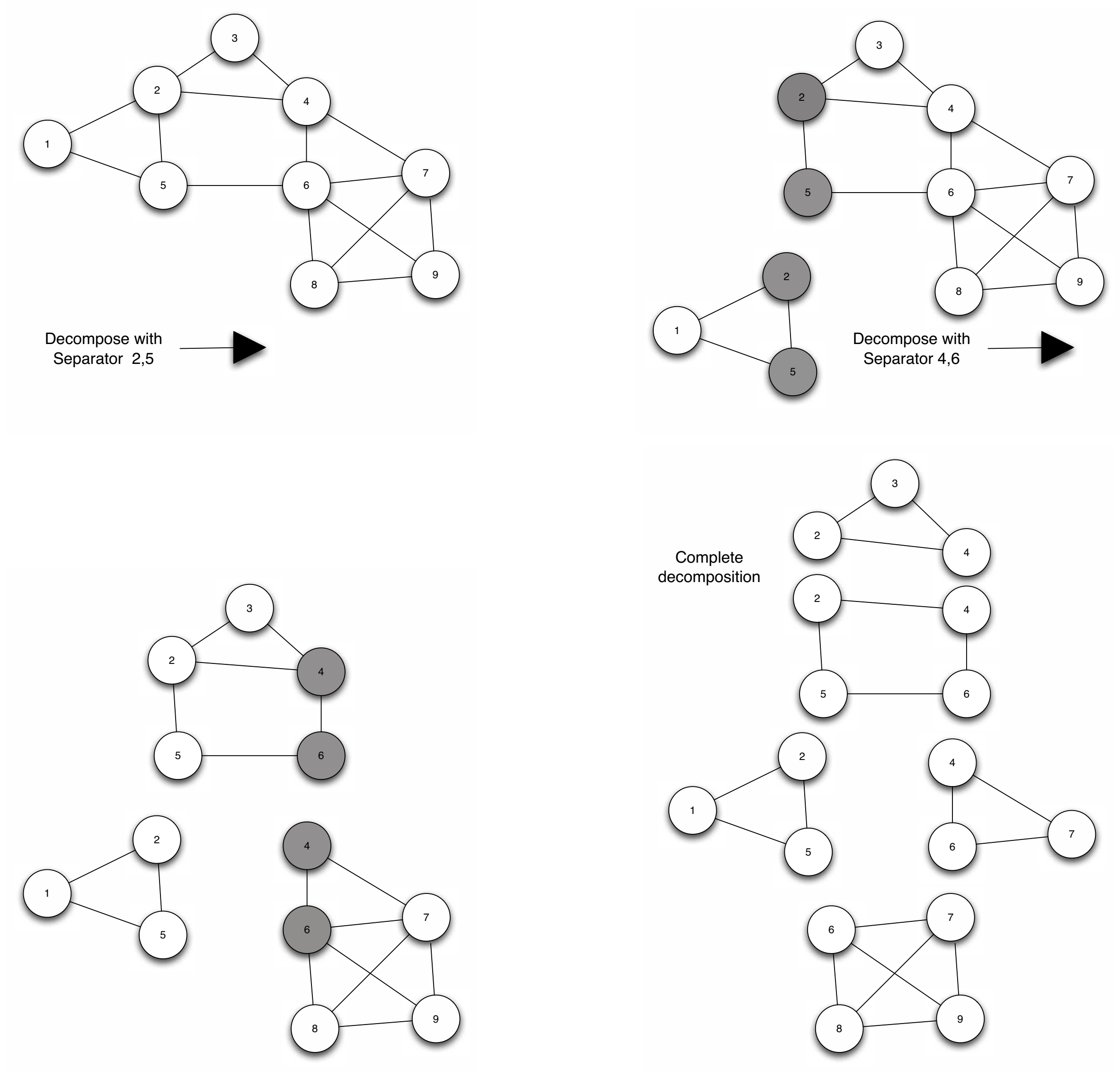


Figure 2.2: Illustration of junction tree decomposition in a non-decomposable graphs. This shows a sequence of iterations of the maximum cardinality search producing the prime components of a graph.

Formally, if X, Y, Z are random variables with joint distribution P, X is *conditionally independent* of Y given Z if for a measurable set A in the sample space of X, $P(X \in A|Y, Z)$ is a function of Z alone. We can write this statement as

$X \perp\!\!\!\perp Y|Z$ (a detailed presentation of conditional independence appears in Dawid, 1980). To establish the connection between graphs and conditional independence ideas the following definitions are needed (Dawid and Lauritzen, 1993). Consider an undirected graph $G = (V, E)$. Associated with each vertex $\alpha \in V$ is a random variable X_α. A distribution P on the vertex set V is said to be $Markov$ with respect to G if, for any decomposition (A, B) of G,

$$A \perp\!\!\!\perp B|A \cap B.$$

Now, let Q and R be the distributions for X_A and X_B, respectively. For the existence of a joint distribution over $A \cup B$ with margins Q and R, Q and R have to be *consistent*. Distributions Q over X_A and R over X_B are said to be *consistent* if they yield the same distribution over $A \cap B$.

Lemma 2.1. *If Q over X_A and R over X_B are consistent, there exists a unique distribution P over $A \cup B$ such that (i) $P_A = Q$; (ii) $P_B = R$; (iii) $A \perp\!\!\!\perp B|A \cap B$.*

Proof. See Dawid and Lauritzen (1993) □

As a direct implication of Lemma 2.1, if P, Q and R have density functions p, q and r, respectively, it is possible to write:

$$\begin{aligned} p(X) &= \frac{q(X_A)r(X_B)}{q(X_{A\cap B})} = \frac{q(X_A)r(X_B)}{r(X_{A\cap B})} \\ &= \frac{p_A(X_A)p_B(X_B)}{p_{A\cap B}(X_{A\cap B})}. \end{aligned}$$

To extend the above representation to a general graph, suppose that $\{P_1, P_2, \ldots, P_k\}$ form a perfect order of prime components of G and $\{p(X_{P_i}) : P \in \mathcal{P}\}$ is a set of consistent marginals for each P_i of graph G. If the distribution of X is Markov with respect to G, for every i such that $P_i \in \mathcal{P}$,

$$P_{i+1} \perp\!\!\!\perp H_i | S_{i+1} \tag{2.1}$$

and the joint density factorizes as (Hammersley and Clifford, 1971):

$$p(X|G) = \frac{\prod_{P \in \mathcal{P}} p(X_P)}{\prod_{S \in \mathcal{S}} p(X_S)}, \tag{2.2}$$

where $\mathcal{S}$ is the set of separators of G. Equations (2.2) and (2.1) are key elements in the analysis of graphical models. All computational efficiencies arise from the decomposition of the sample space of X into subsets of variables based on their graphical relationships, and the ability to model each subset "locally" is a direct consequence of the Markov property of the joint distribution.

2.2 Covariance Selection Models

Graphical structuring of multivariate normal distributions is often referred to as *Gaussian graphical modeling* or covariance selection modeling (Dempster, 1972). If $G = (V, E)$ is an undirected graph and X is a random vector associated with vertices in V, a Gaussian graphical model for X is defined by the assumption that X is normally distributed respecting the conditional independences implied by G. In the normal set up, conditional independence restrictions are simply expressed by zeros in the inverse covariance matrix, or precision matrix. Hence, the canonical parameter $\mathbf{\Omega}$, the precision matrix, belongs to $M(G)$, the set of all positive-definite symmetric matrices with elements equal to zero for all $(i, j) \notin E$. This fact can be formalized by the following theorem:

Theorem 2.1. *Let a $(p \times 1)$ vector $X \sim N(\mu, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ non-singular, $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ be the precision matrix with elements $\omega_{\alpha\beta}$ $(\alpha, \beta = 1, \ldots, p)$ and vertex-set V. For any $\alpha, \beta \in V$*

$$X_\alpha \perp\!\!\!\perp X_\beta | X_{V \setminus \{\alpha,\beta\}} \Longleftrightarrow \omega_{\alpha\beta} = 0. \tag{2.3}$$

Proof. Consider the partition:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{\alpha\beta} & \mathbf{\Sigma}_R \\ \mathbf{\Sigma}'_R & \mathbf{\Sigma}_{V \setminus \{\alpha,\beta\}} \end{pmatrix},$$

where $\mathbf{\Sigma}_{\alpha\beta}$ is a (2×2) covariance matrix for variables α and β, $\mathbf{\Sigma}_R$ is a $(2 \times p-2)$ matrix of covariances between vertices α and β and the remaining vertives. Finally, $\mathbf{\Sigma}_{V \setminus \alpha,\beta}$ is a $(p{-}2{\times}p{-}2)$ covariance matrix for vertices in the set $V \setminus \{\alpha, \beta\}$. From standard linear algebra and normal theory results (Harville, 1997) we have

$$\mathbf{\Omega}_{\alpha\beta} = \mathbf{\Sigma}^{-1}_{\alpha\beta|V \setminus \{\alpha,\beta\}} = \begin{pmatrix} \omega_{\alpha\alpha} & \omega_{\alpha\beta} \\ \omega_{\beta\alpha} & \omega_{\beta\beta} \end{pmatrix}.$$

The covariance matrix of the conditional distribution of $X_{\alpha\beta}|X_{V \setminus \{\alpha,\beta\}}$ is therefore equal to

$$\mathbf{\Sigma}_{\alpha\beta|X_{V \setminus \{\alpha,\beta\}}} = \frac{1}{|\mathbf{\Omega}_{\alpha\beta}|} \begin{pmatrix} \omega_{\beta\beta} & -\omega_{\alpha\beta} \\ -\omega_{\beta\alpha} & \omega_{\alpha\alpha} \end{pmatrix}$$

which implies the result in Equation (2.3). □

Without lost of generality let $\boldsymbol{\mu} = 0$. As in Equation (2.2) the distribution of X is Markov over G and the joint density has the following representation:

$$p(X|\mathbf{\Sigma}_G) = \frac{\prod_{P \in \mathcal{P}} p(X_P|\mathbf{\Sigma}_P)}{\prod_{S \in \mathcal{S}} p(X_S|\mathbf{\Sigma}_S)}, \tag{2.4}$$

where for all prime components and separators, X_P follows a multivariate normal distribution with covariance matrix $\mathbf{\Sigma}_P$. Given G, the joint distribution is completely defined by the component-marginal covariance matrices $\mathbf{\Sigma}_P$, subject to the consistency condition that requires the elements of $\mathbf{\Sigma}_{S_i=P_{i+1}\cap P_i}$ to be common in $\mathbf{\Sigma}_{P_{i+1}}$ and $\mathbf{\Sigma}_{P_i}$, for all $P_i \in \mathcal{P}$. When this holds, $\mathbf{\Omega} \in M(G)$ can be expressed as (Lauritzen, 1996)

$$\mathbf{\Omega} = \sum_{P\in\mathcal{P}} \left[\mathbf{\Sigma}_P^{-1}\right]^0 - \sum_{S\in\mathcal{S}} \left[\mathbf{\Sigma}_S^{-1}\right]^0 \tag{2.5}$$

where K^0 denotes an extension of the matrix K with zeros so as to give it the appropriate dimensions. Equation (2.5) is another example of the local nature of graphical models where the canonical parameter $\mathbf{\Omega}$ of the joint distribution of X is just a function of parameters of the marginal distributions of cliques and separators.

2.3 The Hyper-Inverse Wishart Distribution

In working with covariance selection models, Dawid and Lauritzen (1993) defined a family of Markov probability distributions for covariance matrices on decomposable graphs called the *hyper-inverse Wishart*. If $\mathbf{\Omega} \in M(G)$, the hyper-inverse Wishart for $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$ is denoted by

$$\mathbf{\Sigma} \sim HIW_G(b, \mathbf{D}) \tag{2.6}$$

with degree-of-freedom parameter $b > 0$ and location matrix $\mathbf{D} > 0$. This distribution is the unique hyper-Markov distribution for $\mathbf{\Sigma}$ with consistent clique-marginals that are inverse Wishart distributed. The joint density decomposes as

$$p(\mathbf{\Sigma}|b, \mathbf{D}) = \frac{\prod_{C\in\mathcal{C}} p(\mathbf{\Sigma}_C|b, \mathbf{D}_C)}{\prod_{S\in\mathcal{S}} p(\mathbf{\Sigma}_S|b, \mathbf{D}_S)} \tag{2.7}$$

where for each $C \in \mathcal{C}$, $\mathbf{\Sigma}_C \sim IW(b, \mathbf{D}_C)$ with density

$$p(\mathbf{\Sigma}_C|b, \mathbf{D}_C) = \frac{|\frac{\mathbf{D}_C}{2}|^{\left(\frac{b+|C|-1}{2}\right)}}{\Gamma_{|C|}\left(\frac{b+|C|-1}{2}\right)}|\mathbf{\Sigma}_C|^{-(b+2|C|)/2} exp\left(-\frac{1}{2}tr(\mathbf{\Sigma}_C^{-1}\mathbf{D}_C)\right). \tag{2.8}$$

$\mathbf{D}_C$ is the positive-definite symmetric diagonal block of $\mathbf{D}$ corresponding to $\mathbf{\Sigma}_C$ and $\Gamma_k(a)$ is the multivariate gamma function:

$$\Gamma_k(a) = \pi^{\frac{k(k-1)}{4}} \prod_{i=0}^{i=k-1} \Gamma(a - \frac{i}{2}).$$

In the case of a decomposable model, the expected value of $\mathbf{\Omega}$ is available in closed form. Using the representation in (2.5) and given $\mathbf{\Sigma} \sim HIW_G(b, \mathbf{D})$, the expected value of $\mathbf{\Omega}$ takes the form:

$$\begin{aligned} E(\mathbf{\Omega}|b, D) &= \sum_{C \in \mathcal{C}} [E(\mathbf{\Omega}_C|b, D_C)]^0 - \sum_{S \in \mathcal{S}} [E(\mathbf{\Omega}_S|b, D_S)]^0 \qquad (2.9) \\ &= \sum_{C \in \mathcal{C}} \left[(b + |C| - 1)(\mathbf{D}_C)^{-1}\right]^0 - \sum_{S \in \mathcal{S}} \left[(b + |S| - 1)(\mathbf{D}_S)^{-1}\right]^0. \end{aligned}$$

Decomposable graphs consist entirely of complete prime components which means that Equations (2.7) and (2.8) are enough to express the density of $\mathbf{\Sigma}$. The tractability of these models is explained by the fact that, while the graphical structure determines which elements of the covariance matrix appear in the density, the elements that do appear are unconstrained in the sense that there are no constraints on the clique (local) level. This is not the case for non-decomposable models where the presence of non-complete prime components impose additional local constraints. To deal with non-decomposable graphs, a consistent distribution analogous to (2.8) that incorporates these local constraints is needed. Grone *et al.*

(1984) showed that when considering an incomplete covariance matrix (where only the entries corresponding to edges or on the diagonal are filled in), if the matrix can be completed to be a positive definite matrix consistent with the graph, this completion is unique. Taking advantage of this fact, Roverato (2002) is able to generalize the hyper-inverse Wishart to incorporate the constraints of non-complete prime components and define a density for $\mathbf{\Sigma}_P$ as a function of its positive definite completion. Let the free elements be determined by the edge set $\mathcal{E}$, so we give the density argument as $\mathbf{\Sigma}_P^{\mathcal{E}}$. The expression for the density is:

$$p(\mathbf{\Sigma}_P^{\mathcal{E}}|G, b, \mathbf{D}) \propto |\mathbf{\Sigma}_P|^{-\left(\frac{b-2}{2}\right)} J(\mathbf{\Omega}_P^{\mathcal{E}} \to \mathbf{\Sigma}_P^{\mathcal{E}}) \exp\left(-\frac{1}{2}\mathbf{\Sigma}_P^{-1}\mathbf{D}_P\right) \tag{2.10}$$

where $\mathbf{\Sigma}$ is the positive definite completion of $\mathbf{\Sigma}^{\mathcal{E}}$ and $J(\mathbf{\Omega}_P^{\mathcal{E}} \to \mathbf{\Sigma}_P^{\mathcal{E}})$ is the Jacobian of the transformation from $\mathbf{\Omega}_P^{\mathcal{E}}$ (which has zeroes for off-diagonal entries not corresponding to edges in $\mathcal{E}$) to $\mathbf{\Sigma}_P^{\mathcal{E}}$. This density is obtained from a Wishart prior on $\mathbf{\Omega}_P$, conditioned on $\mathbf{\Omega}_P$ consistent with G, by a change of variables. This extension is consistent with the definition of the hyper-inverse Wishart and generalizes the unique hyper-Markov distribution for any graph G where the joint density decomposes as in (2.7).

Grone has also shown that if the sub-matrices corresponding to the cliques in a decomposable graph are positive definite, then a positive definite completion consistent with the graph always exists. This is reflected in the density for decomposable graphs where none of the "non-free" elements appear in either (2.7) or (2.8), hence not affecting the density at all. Combined with the positive-definite completion of non-complete prime components, this establish a way to evaluate the "non-free" elements of $\mathbf{\Sigma}$ as a function of its free elements. This is done via the completion operation described in Lauritzen (1996) and in a general context in Massam and Neher (1998); that is, given the perfect ordering of prime components and separators, and defining $A_{i-1} = H_{i-1} \setminus S_i$ for each i, the "non-free" elements are directly computed by

$$\mathbf{\Sigma}_{R_i,A_{i-1}} = \mathbf{\Sigma}_{R_i,S_i}\mathbf{\Sigma}_{S_i}^{-1}\mathbf{\Sigma}_{S_i,A_{i-1}}. \tag{2.11}$$

2.4 Prior and Posterior for Covariance Matrices

From a Bayesian perspective, conditional on a graph G, inference on the covariance matrix $\mathbf{\Sigma}$ in a Gaussian graphical model are based on the posterior

$$p(\mathbf{\Sigma}|X,G) \propto p(X|\mathbf{\Sigma},G)p(\mathbf{\Sigma}|G)$$

where the likelihood $p(X|\mathbf{\Sigma},G)$ is define in (2.4) and the prior $p(\mathbf{\Sigma}|G)$ represents all the relevant information about $\mathbf{\Sigma}$ respecting the restriction imposed by G.

Giudici (1996) discusses the major approaches to prior specification for $\mathbf{\Sigma}$, comparing the "local priors" described in Dawid and Lauritzen (1993), and the "global priors" based on the conditional approach in Dickey (1971). These priors have the desirable property that $p(\mathbf{\Sigma}|G)$ is consistent over graphs: the (i,j) element of $\mathbf{\Omega}$ has the same prior whenever the graph does not constrain the (i,j) element to be zero.

The hyper-inverse Wishart turns out to be the conjugate "local prior" for any graph G; so, if $p(\mathbf{\Sigma}|G) = HIW_G(b,\mathbf{D})$, for a random sample of size n and $\mathbf{X} = \{X^{(1)},\ldots,X^{(n)}\}$, the posterior for $\mathbf{\Sigma}$ is $HIW_G(b+n,\mathbf{D}+\mathbf{S}_X)$ where $\mathbf{S}_X$ is the cross product matrix $\mathbf{X}\mathbf{X}'$. This is a direct result of the decomposition of both the prior and the likelihood (as in 2.7 and 2.4) together with standard conjugacy of normals and inverse-Wisharts giving the following posterior:

$$\begin{aligned} p(\mathbf{\Sigma}|\mathbf{X},G) &= \frac{\prod_{P\in\mathcal{P}} p(X_P|\mathbf{\Sigma}_P)}{\prod_{S\in\mathcal{S}} p(X_S|\mathbf{\Sigma}_S)} \frac{\prod_{P\in\mathcal{P}} p(\mathbf{\Sigma}_P|b,\mathbf{D}_P)}{\prod_{S\in\mathcal{S}} p(\mathbf{\Sigma}_S|b,\mathbf{D}_S)} \\ &= \frac{\prod_{P\in\mathcal{P}} p(X_P|\mathbf{\Sigma}_P)p(\mathbf{\Sigma}_P|b,\mathbf{D}_P)}{\prod_{S\in\mathcal{S}} p(X_S|\mathbf{\Sigma}_S)p(\mathbf{\Sigma}_S|b,\mathbf{D}_S)} \end{aligned}$$

where the prior for each prime component (and separator) is an inverse Wishart. Hence the posterior for every $P \in \mathcal{P}$ can be written as:

$$\begin{aligned} p(\mathbf{\Sigma}_P|\mathbf{X},G) &= p(X_P|\mathbf{\Sigma}_P)p(\mathbf{\Sigma}_P|b,\mathbf{D}_P) \\ &\propto |\mathbf{\Sigma}_P|^{-(b+2|P|)/2} exp\left(-\frac{1}{2}tr(\mathbf{\Sigma}_P^{-1}\mathbf{D}_P)\right)|\mathbf{\Sigma}_P|^{-n/2} exp\left(-\frac{1}{2}tr(\mathbf{\Sigma}_P^{-1}\mathbf{S}_{X_P})\right) \\ &\propto |\mathbf{\Sigma}_P|^{-(n+b+2|P|/2)} exp\left(-\frac{1}{2}tr\left[\mathbf{\Sigma}_P^{-1}(\mathbf{S}_{X_P}+\mathbf{D}_P)\right]\right), \qquad (2.12) \end{aligned}$$

that is

$$(\mathbf{\Sigma}_P|\mathbf{X},G) \sim IW(b+n, \mathbf{D}_P + \mathbf{S}_{X_P}). \qquad (2.13)$$

Marginal distributions for $\mathbf{X}$ (Dawid and Lauritzen, 1993), take the form of *hyper-t distributions*. For a clique C the marginal distribution $p(X_C|G)$ is a matrix t distribution denoted by $T_b(0,\mathbf{D}_C,I)$ with density defined in Appendix C. The overall marginal distribution is Markov and is defined in the same way as (2.2) and denoted by $HT_G(0,\mathbf{D},b)$.

Chapter 3

Model Determination in High-Dimensional Graphical Models

So far, all aspects of Gaussian graphical models presented rely on the knowledge of the graph G defining the structure of conditional independence in the covariance matrix. More often than not the graph is not known and estimating covariances matrices will involve inferences about the graph as well. This is in fact one of the hardest challenges faced by the Gaussian graphical model literature. Even when the dimension $|V| = p$ is of moderate size, it is essentially impossible to enumerate and compare the relevance of all $2^{\frac{p(p-1)}{2}}$ possible graphs.

The development of stochastic computational tools to search the space of graphs is a fundamental step to enable the analysis of high-dimensional problems. In the Bayesian framework, MCMC methods are a natural way to access the posterior distribution of models and, when the number of variables is relatively small, these methods are capable of efficiently identifying graphs with high posterior mass. However, in high-dimensional problems these approaches are slow to converge and tend not to provide a comprehensive summary of the model space.

This opens the door for the development of search algorithms that can rapidly identify relevant areas of the model space and explore the neighborhood around it. Sparsity is a central guiding principal here: it is of interest to develop parsimonious models - models of the lowest dimensional-parameters capable of adequately represent observed data configurations, especially in higher dimensional distributions and when in the "large p small n" paradigm.

A number of recent papers have addressed the question of improving computational methods for Gaussian graphical model determination. Due to computational efficiencies and accessible distribution theory, the key focus of the literature has been on decomposable graphs. Giudici and Green (1999) describe a efficient way to implement a reversible jump MCMC in the space of decomposable graphs while Wong *et al.* (2003) show how the same efficient tools can be incorporated in a much simpler Gibbs sampler. This general line of development has recently been extended beyond decomposable models in works by Roverato (2002), Dellaportas *et al.* (2003) and Atay-Kayis and Massam (2005).

Unlike the recent literature that basically deals with very small examples, this Chapter tries to understand the performance and scalability of search methods as dimension grows. First, I describe two MCMC methods for model space exploration to serve as benchmarks for a novel, parallelizable, stochastic search algorithm that can quickly traverse high-dimensional model spaces. These methods are tested in examples that combine simple simulated datasets of moderate size ($p = 12$ and $p = 15$) with larger scale ($p = 150$) data analysis motivated by gene expression studies. Another important aspect of this chapter is that both decomposable and non-decomposable models are considered.

3.1 Marginal Likelihood for Graphs

From a Bayesian perspective, model selection involves the exploration of the posterior distribution of graphs, given by:

$$p(G|X) \propto p(X|G)p(G) \tag{3.1}$$

where $p(X|G)$ is the *marginal likelihood* of G and $p(G)$ represents its prior.

In a Gaussian graphical model where $|V| = p$ and $X = \{X_1, \ldots, X_n\}$, the marginal likelihood for any graph G is given by the following integral

$$p(X|G) = \int_{\mathbf{\Sigma}^{-1}=\mathbf{\Omega}\in M(G)} p(X|\mathbf{\Sigma}, G)p(\mathbf{\Sigma}|G)d\mathbf{\Sigma} \tag{3.2}$$

where, as before, $M(G)$ defines the set of all positive-definite symmetric matrices respecting the constraints of G.

Using hyper-inverse Wishart priors for $\mathbf{\Sigma}$, the computation of the marginal likelihood for a decomposable graph is straightforward. The prior normalizing constant and a factor of $(2\pi)^{-np/2}$ from the likelihood can be pulled through the integral, and $p(X|G)$ becomes a simple function of the prior and posterior normalizing constants, $h(G, b, \mathbf{D})$ and $h(G, b^*, \mathbf{D}^*)$:

$$p(X|G) = (2\pi)^{-np/2}\frac{h(G, b, \mathbf{D})}{h(G, b^*, \mathbf{D}^*)} \tag{3.3}$$

where $b^* = b + n$ and $\mathbf{D}^* = \mathbf{D} + \mathbf{S}_X$.

The combination of Equations (2.7) and (2.8) shows that the normalizing constant of a hyper-inverse Wishart is a function of normalizing constants of the corresponding inverse Wisharts of cliques and separators, namely:

$$h(G,b,\mathbf{D}) = \frac{\prod_{C\in\mathcal{C}} |\frac{\mathbf{D}_C}{2}|^{\left(\frac{b+|C|-1}{2}\right)} \Gamma_{|C|}\left(\frac{b+|C|-1}{2}\right)^{-1}}{\prod_{S\in\mathcal{S}} |\frac{\mathbf{D}_S}{2}|^{\left(\frac{b+|S|-1}{2}\right)} \Gamma_{|S|}\left(\frac{b+|S|-1}{2}\right)^{-1}}. \tag{3.4}$$

In the non-decomposable case the marginal likelihood for G can still be expressed by (3.3) but, now, at least one incomplete prime component is involved in the computation of $h(G,b,\mathbf{D})$. As seen in (2.10), the density function for a non-complete prime component is only known up to the normalizing constant; the integration of a Wishart kernel over a set of constraints positive definite matrices has no close form solution and numerical approximations are necessary. A Monte Carlo method to compute marginal likelihoods for non-decomposable graphs is presented next.

3.1.1 Computing Marginal Likelihoods for Non-Decomposable Models

In order to facilitate the discussion, assume that one incomplete prime component P constitutes the entire graph, i.e. $G = P = (V, E)$, so the marginal likelihood is simply a function of

$$h(P,b,\mathbf{D}) = \int_{\mathbf{\Omega}_P\in M(P)} |\mathbf{\Sigma}_P|^{-\left(\frac{b-2}{2}\right)} J(\mathbf{\Omega}_P^{\mathcal{E}} \to \mathbf{\Sigma}_P^{\mathcal{E}}) \exp\left\{-\frac{1}{2}\mathbf{\Sigma}_P^{-1}\mathbf{D}_P\right\} d\mathbf{\Omega}_P. \tag{3.5}$$

As mentioned before, no analytical solution is available to solve for the evaluation of the normalizing constant in (3.5). Atay-Kayis and Massam (2005) exploit two transformations that generalize properties of the Bartlett decomposition (Bartlett, 1933) to restricted Wishart matrices, defining a way to generate samples from $\mathbf{\Omega} \in M(G)$ that can be used to compute a Monte Carlo estimate of $h(P,b,\mathbf{D})$. An

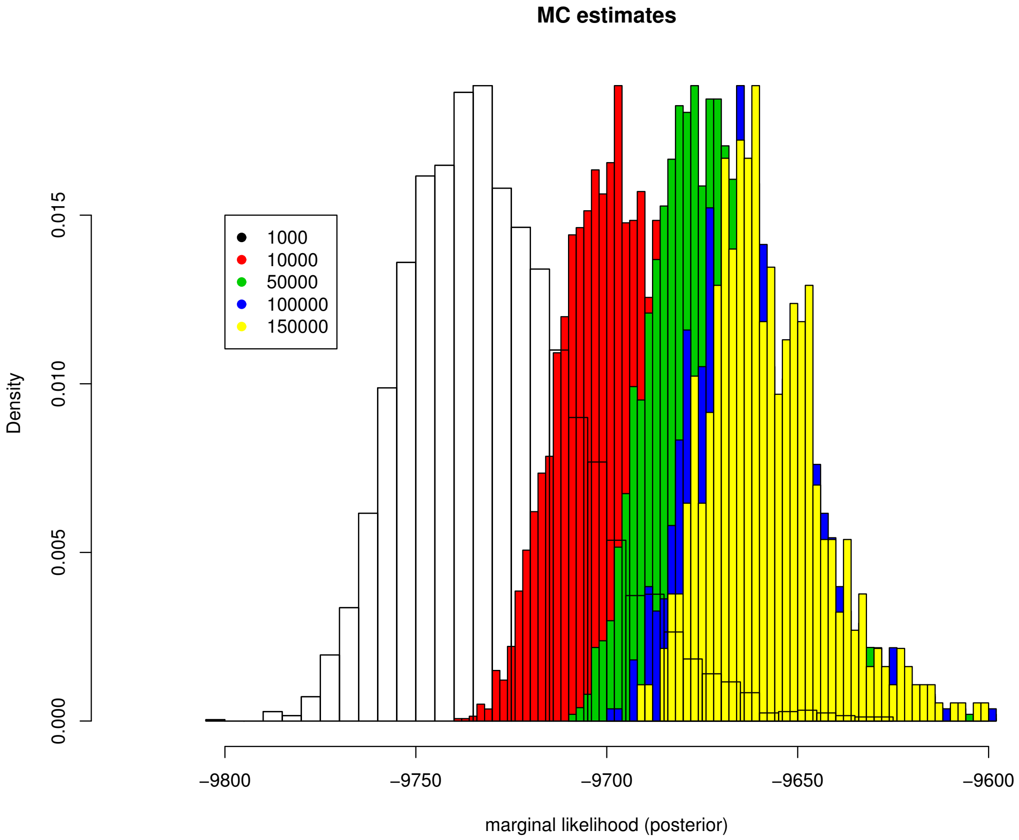


Figure 3.1: Approximations of the marginal likelihood for G (non-decomposable). Each histogram represents multiple values of the approximation based on different number of MC samples.

essential fact in this development is that for any matrix $\mathbf{x} \in M(G)$ the Cholesky decomposition $\boldsymbol{\phi}$ has "free" elements ϕ_{ij}, $(i, j) \in E$ (Roverato, 2002), and remaining elements ϕ_{ij}, $(i, j) \notin E$ given by a direct function of the free elements. This result implies that the free elements of $\boldsymbol{\phi}$ are enough to define $\boldsymbol{\Omega}$ and the integration in (3.5) can be written as a explicit function of $\boldsymbol{\phi}$, simplifying the desired approximation. Being more specific to the problem in hand, let $\boldsymbol{\Omega} = \boldsymbol{\phi}'\boldsymbol{\phi}$ and $\boldsymbol{\psi} = \boldsymbol{\phi}\mathbf{T}^{-1}$ where $\mathbf{D}^{-1} = \mathbf{T}'\mathbf{T}$ where both $\boldsymbol{\phi}$ and $\mathbf{T}$ are upper triangular matrices.

After computing the Jacobians $J(\mathbf{\Omega} \to \boldsymbol{\phi})$ and $J(\boldsymbol{\phi} \to \boldsymbol{\psi})$ it can be shown that the free elements ψ_{ij}, $(i,j) \in E$, are independent normal $(i \neq j)$ and square roots of χ^2 $(i = j)$ random variables; The integral in (3.5) can then be written as a expectation, given by:

$$\begin{aligned} h(P, b, \mathbf{D}) &= C_b \int exp\left(-\frac{1}{2} \sum_{(i,j) \notin E} \psi_{ij}^2 \right) && (3.6) \\ &\times \prod_{i=1}^{|V|} \Gamma\left(b + \nu_i\right)^{-1} \left(\frac{\psi_{ii}^2}{2} \right)^{\frac{b+\nu_i-2}{2}} exp\left[-\frac{1}{2} \psi_{ii}^2 \right] \prod_{(i,j) \in E} \frac{1}{\sqrt{2\pi}} exp\left[-\frac{1}{2} \psi_{ij}^2 \right] d\boldsymbol{\psi} \\ &= C_b \mathbf{E} \left[exp\left(-\frac{1}{2} \sum_{(i,j) \notin E} \psi_{ij}^2 \right) \right] && (3.7) \end{aligned}$$

where C_b is a constant

$$C_b = \left(\prod_{i=1}^{|V|} 2^{\frac{b+\nu_i}{2}} (2\pi)^{\frac{\nu_i}{2}} \Gamma\left(\frac{b + \nu_i}{2} \right) t_{ii}^{\frac{b+z_i-1}{2}} \right)$$

with ν_i being the number of neighbors of vertex i subsequent to it in order of the vertices, and z_i is equal to the total number of neighbors of i plus 1. The form of (3.6) shows that the expectation is taken with respect to the distribution with density equal to the product of independent χ^2 with $b + \nu_i$ degrees of freedom and standard normal distributions. Approximating (3.5) via Monte Carlo is simply done by sampling ψ_{ij} for all $(i,j) \in E$ from normals and χ^2 distributions, and computing:

$$\hat{h}(P, b, \mathbf{D}) \approx \frac{1}{N} \sum_{l=1}^{N} exp\left(-\frac{1}{2} \sum_{(i,j) \notin E} {\psi_{ij}^{(l)}}^2 \right) \tag{3.8}$$

where each ψ_{ij} for $(i,j) \notin E$ is equal to:

- if $i = 1$,

$$\psi_{ij} = -\sum_{k=1}^{j-1} \psi_{ik} t_{\langle kj]},$$

- and for $i > 1$,

$$\psi_{ij} = \sum_{k=i}^{j-1} \psi_{ik} t_{\langle kj]} - \sum_{r=1}^{i-1} \left(\frac{\psi_{ri} + \sum_{l=r}^{i-1} \psi_{rl} t_{\langle li]}}{\psi_{ii}} \right) \left(\psi_{rj} + \sum_{l=r}^{j-1} \psi_{rl} t_{\langle kj]} \right).$$

with $t_{\langle ij]} = t_{ij}/t_{jj}$.

Alternatives to Massam's method include works by Roverato (2002) and Dellaportas *et al.* (2003). In Roverato (2002) an importance sampling method to compute (3.5) is developed. The method uses an approximating decomposable model, with edge set E^* containing E to create a tractable importance function that is used to generate samples of $\boldsymbol{\phi}$. Similarly, Dellaportas *et al.* (2003) use a change of variables and write the normalizing constant as an expectation over the transformed space. The expectation is then estimated through an importance function based on samples from a multivariate normal random variables.

Due to its simplicity and lack of "extra" approximation steps, the method of Atay-Kayis and Massam (2005) is used in all examples involving the calculation of marginal likelihoods for non-decomposable graphs and will also be useful in Chapter 4 where methods to generate samples directly from a HIW_G given a graph G are discussed.

It is important to point out that scalability is an issue here. In a large non-decomposable graph with many incomplete prime components, the above approximation has to be carried out many times so introducing variation in the marginal likelihood estimate. An illustration of this problem is showed in Figure 3.1 where for a non-decomposable graph with 150 nodes, empirical distributions of the approximated marginal likelihood are displayed. This is an indication that many Monte Carlo samples might be needed before a stable estimator is available. A detailed discussion of this problem is presented later on this chapter, in the context of graphical model selection.

3.2 Priors over Graphs

A uniform prior over all graphs assigns most of its mass to graphs with half of the total number of possible edges. The number of possible edges in a graph with p nodes is $T = p(p-1)/2$ and so, for large p, the uniform prior favors models where the number of edges is quite large. To illustrate this concept, simulations from the uniform prior on the space of decomposable models are displayed in Figure 3.2. This indicates that the average number of edges explodes very quickly as the number of nodes increase.

In many situations, including the examples in this dissertation, parsimonious representations of the conditional independence structure are of interest and sparsity encouraging priors are needed. A Bernoulli prior on each edge inclusion probability with parameter β is used here as an initial "sparsity inducing" model; a graph with e edges has prior probability proportional to $\beta^e(1-\beta)^{(T-e)}$. This distribution has its peak at $T \times \beta$ edges for an unrestricted p node graph providing a direct way to control the prior complexity of the model. Figure 3.3 displays histograms for number of edges in graphs sampled from this prior when $\beta = 2/(T-1)$. It is clear that the explosive behavior of the uniform prior is no longer present in the Bernoulli prior case.

This approach to prior specification penalizes the number of edges, with the

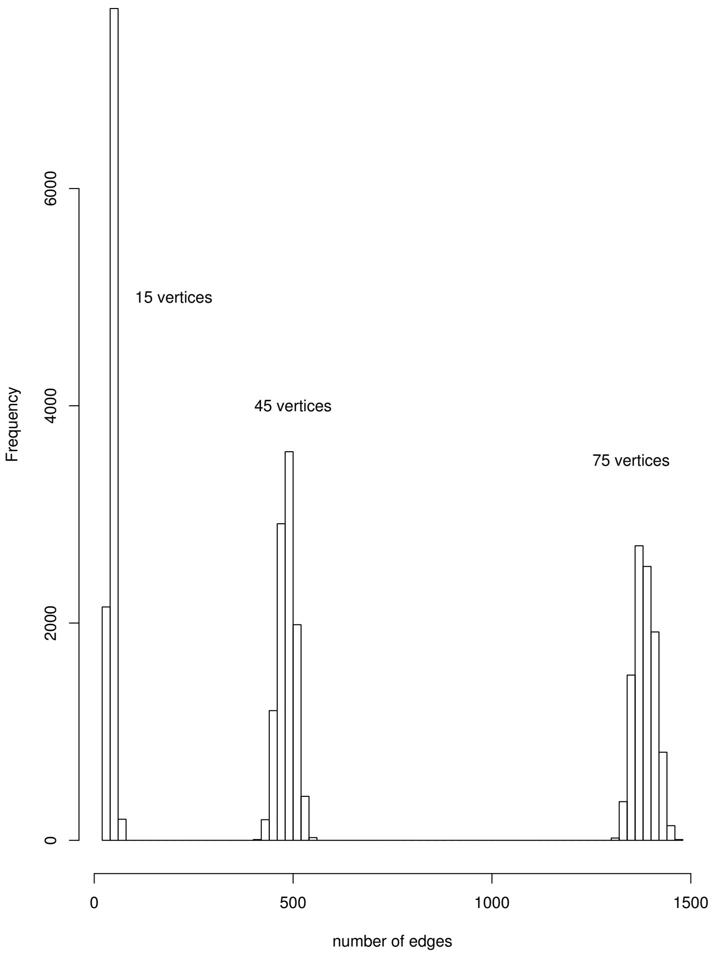


Figure 3.2: For each different number of vertices listed, the histogram represents the prior mass on different numbers of edges under a prior that is uniform over decomposable graphs. The histogram is based on sampling from the prior with a Metropolis-Hastings algorithm.

view that, if choosing between two edges, preference is given to the edge resulting in the greatest increase of the graph's marginal likelihood regardless of the rest of the graph's structure. One could, of course, penalize other measures of complexity such as the maximum or average prime component size, number of cliques, etc.

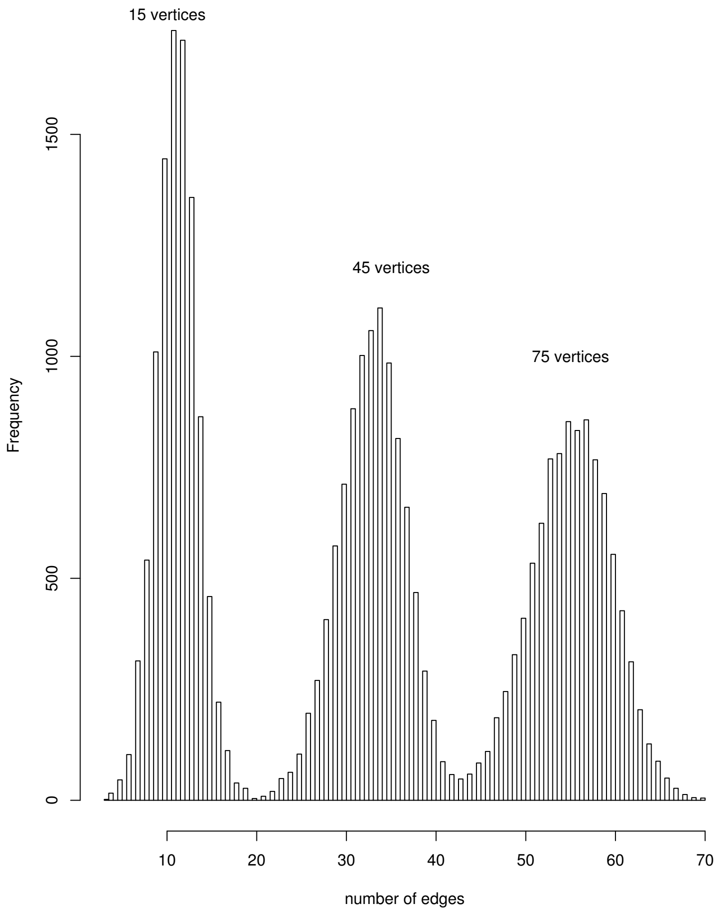


Figure 3.3: For each different number of vertives listed, the histogram represents the prior mass on different numbers of edges where consideration is restricted to decomposable graphs. The prior mass of a graph is proportional to $\beta^E(1-\beta)^{T-E}$, where $\beta = 2/(p-1)$. The histogram is based on sampling from the prior with a Metropolis-Hastings algorithm.

Wong *et al.* (2003) developed an approach that equalizes the prior probability of graphs with different numbers of edges; for decomposable graphs, this requires estimating the fraction of the total number of decomposable graphs with each number of edges which can, in larger problems, be computationally intensive.

3.3 Local MCMC Updates in Decomposable Models

In all model search methods considered in this work, moves in graphical model space are local; that is to say, moves from graph G to graph G' differ only by adding or deleting one edge. Once again, decomposable graphs present a very important computational advantage; computing the marginal likelihood ratio $p(X|G)/p(X|G')$ is facilitated by the similarity of G and G' that allow local updates of the ratio involving at most two cliques and one separator. The following theorem helps to clarify this notion.

Theorem 3.1. *Suppose that $G = (V, E)$ and $G' = (V, E')$ are decomposable graphs differing by one edge (i, j), $E \setminus E' = (i, j)$. Let $\{C_1, S_2, C_2, S_3 \ldots, C_k, S_k\}$ be the perfect order of cliques and separators of G. Then:*

(i) *The edge (i, j) is contained in a single clique of G;*

(ii) *If $(i, j) \in C_q$ then either $i \notin S_q$ or $j \notin S_q$;*

(iii) *If $j \notin S_q$ and $C_{q_1} = C_q \backslash \{j\}$ and $C_{q_2} = C_q \backslash \{i\}$ then $S_{q_1} = S_q$,$S_{q_2} = C_q \backslash \{i, j\}$ and $\{C_1, S_2, \ldots, C_{q-1}, S_{q-1}, C_{q_1}, S_q, C_{q_2}, S_{q_2}, \ldots, C_k, S_k\}$ form a perfect order of cliques and separators of G'.*

Proof. See Theorem 1 of Giudici and Green (1999), and Lemma 2.20 of Lauritzen (1996) □

Combined with the factorization of the marginal likelihood implied by (3.4) Theorem 3.1 means that computing the likelihood ratio between G and G' results in cancellation of all terms except those involving $\{C_q, C_{q_1}, C_{q_2}, S_{q_2}\}$. Write the “prior” part of the marginal likelihood ratio as

$$\begin{aligned}
\frac{h(G,b,\mathbf{D})}{h(G',c,\mathbf{D})} &= \frac{\left|\frac{\mathbf{D}_{C_q}}{2}\right|^{\left(\frac{b+|C_q|-1}{2}\right)} \Gamma_{|C_q|}\left(\frac{b+|C_q|-1}{2}\right)^{-1} \left|\frac{\mathbf{D}_{S_{q_2}}}{2}\right|^{\left(\frac{b+|S_{q_2}|-1}{2}\right)} \Gamma_{|S_{q_2}|}\left(\frac{b+|S_{q_2}|-1}{2}\right)^{-1}}{\left|\frac{\mathbf{D}_{C_{q_1}}}{2}\right|^{\left(\frac{b+|C_{q_1}|-1}{2}\right)} \Gamma_{|C_{q_1}|}\left(\frac{b+|C_{q_1}|-1}{2}\right)^{-1} \left|\frac{\mathbf{D}_{C_{q_2}}}{2}\right|^{\left(\frac{b+|C_{q_2}|-1}{2}\right)} \Gamma_{|C_{q_2}|}\left(\frac{b+|C_{q_2}|-1}{2}\right)^{-1}} \\
&= \frac{\left|\mathbf{D}_{C_q}\right|^{\left(\frac{b+|C_q|-1}{2}\right)} \left|\mathbf{D}_{S_{q_2}}\right|^{\left(\frac{b+|S_{q_2}|-1}{2}\right)} \Gamma_{|S_{q_2}|}\left(\frac{b+|S_{q_2}|}{2}\right)}{\left|\mathbf{D}_{C_{q_1}}\right|^{\left(\frac{b+|C_{q_1}|-1}{2}\right)} \left|\mathbf{D}_{C_{q_2}}\right|^{\left(\frac{b+|C_{q_2}|-1}{2}\right)} \Gamma_{|S_{q_2}|}\left(\frac{b+|S_{q_2}|+1}{2}\right) 2\sqrt{\pi}},
\end{aligned} \tag{3.9}$$

where the final simplification result from properties of the multivariate Γ-function (Muirhead, 1982). Similar expressions can be derived for the “posterior” part of the ratio $h(G', b^*, \mathbf{D}^*)/h(G, b^*, \mathbf{D}^*)$ and when combined with (3.9) the desired marginal likelihood ratio is obtained.

Wong *et al.* (2003) show that (3.9) can be simplified even further and that the determinants needed for the likelihood ratio can be computed using only the Cholesky decomposition of $\mathbf{D}_{C_q}$ and $\mathbf{D}^*_{C_q}$. By partitioning $\mathbf{D}_{C_q}$ as

$$\mathbf{D}_{C_q} = \begin{pmatrix} \mathbf{D}_{S_{q_2}} & \mathbf{D}_{S_{q_2}Q} \\ \mathbf{D}_{QS_{q_2}} & \mathbf{D}_Q \end{pmatrix}$$

where $Q = \{i, j\}$, and

$$\mathbf{D}_Q = \begin{pmatrix} d_{ii} & d_{ij} \\ d_{ji} & d_{jj} \end{pmatrix},$$

Equation (3.9) can then be written as a function of

$$\begin{aligned}
\mathbf{D}_{Q.S_{q_2}} &= \mathbf{D}_Q - \mathbf{D}_{QS_{q_2}}(\mathbf{D}_{S_{q_2}S_{q_2}})^{-1}\mathbf{D}_{QS_{q_2}}, \\
d_{ii.S_{q_2}} &= d_{ii} - \mathbf{D}_{iS_{q_2}}(\mathbf{D}_{S_{q_2}})^{-1}\mathbf{D}_{S_{q_2}i}, \\
d_{jj.S_{q_2}} &= d_{jj} - \mathbf{D}_{jS_{q_2}}(\mathbf{D}_{S_{q_2}})^{-1}\mathbf{D}_{S_{q_2}j}
\end{aligned}$$

and corresponding quantities for $\mathbf{D}^*$. This is true by noting that (Theorem 13.3.8 Harville, 1997)

$$\begin{aligned}\left|\mathbf{D}_{C_q}\right| &= \left|\mathbf{D}_{Q.S_{q2}}\right|\left|\mathbf{D}_{S_{q2}}\right|,\\ \left|\mathbf{D}_{C_{q1}}\right| &= d_{ii.S_{q2}}\left|\mathbf{D}_{S_{q2}}\right|,\end{aligned}$$

and

$$\left|\mathbf{D}_{C_{q2}}\right| = d_{jj.S_{q2}}\left|\mathbf{D}_{S_{q2}}\right|.$$

After substituting the appropriate terms, (3.9) takes the form:

$$\frac{h(G,b,\mathbf{D})}{h(G',c,\mathbf{D})} = \frac{\left|\mathbf{D}_{Q.S_{q2}}\right|^{\left(\frac{b+|C_q|-1}{2}\right)}\Gamma_{|S_{q2}|}\left(\frac{b+|S_{q2}|}{2}\right)}{\left(d_{ii.S_{q2}}\right)^{\left(\frac{b+|C_{q1}|-1}{2}\right)}\left(d_{jj.S_{q2}}\right)^{\left(\frac{b+|C_{q2}|-1}{2}\right)}\Gamma_{|S_{q2}|}\left(\frac{b+|S_{q2}|+1}{2}\right)2\sqrt{\pi}}. \tag{3.10}$$

Now, consider the Cholesky decomposition of $\mathbf{D}_{C_q} = L'L$, partitioned as

$$L' = \begin{pmatrix} L_{S_{q2}} & 0 \\ L_{QS_{q2}} & L_Q \end{pmatrix}$$

where

$$L_Q = \begin{pmatrix} l_{ii} & 0 \\ l_{ji} & l_{jj} \end{pmatrix}.$$

It is straightforward to show that

$$\begin{aligned}\left|\mathbf{D}_{Q.S_{q2}}\right| &= l_{ii}^2 l_{ll}^2,\\ d_{ii.S_{q2}} &= (l_{ii})^2,\end{aligned}$$

and

$$d_{jj.S_{q2}} = (l_{ji})^2 + (l_{jj})^2,$$

giving all the necessary quantities to compute the marginal likelihood ratio.

In contrast, when non-decomposable graphs are considered, there is no guarantee that significant cancellations in the marginal likelihood ratio between "neighboring" graphs is available. While the likelihoods still factor over prime components, a single edge change may radically alter the junction tree. As an example, imagine starting with a graph where all the nodes are connected in a chain, and then adding the edge that completes the full cycle. The single edge change takes us from a decomposable graph, with $p-1$ prime components, to a non-decomposable graph with a single prime component; in this case, there are no cancellations in the computation of the marginal likelihood ratio.

3.4 Markov Chain Monte Carlo Algorithms

MCMC tools are frequently used for exploring the space of graphical structures (e.g. Dellaportas and Forster, 1999; Giudici and Castelo, 2003; Giudici and Green, 1999; Madigan and York, 1995). These methods simulate Markov chains that theoretically converge to the posterior distribution of models, say $p(G|X)$. In graphical models, for problems with even a moderate number of variables, the model space is so large that the chain will typically not converge in any practical sense, so MCMC methods can only be realistically used as stochastic search procedures to identify models with high posterior probability.

This section focus on efficient MCMC methods to explore the space of graphs that integrate out parameters and work directly with the marginal likelihood of models. This way there is no need to sample from the full conditional of parameters and the use of a reversible jump methodology is avoided.

Generally, MCMC model search approaches can be classified as Gibbs sampling

(Gelfand and Smith, 1990; George and McCulloch, 1993) and Metropolis-Hasting algorithms. Similarly to the method developed for variable selection (George and McCulloch, 1993), the fixed scan Gibbs sampler for Gaussian graphical models updates one edge at a time according to its full conditional distribution. Let the random variable e_{ij} represent the edge set E for the current graph $G = (V, E)$ in the MCMC, where for all $i < j$, $e_{ij} = 1$ if $(i, j) \in E$ and $e_{-ij} = E \setminus (i, j)$. Each edge can then be updated by sampling from

$$p(e_{ij}|e_{-ij}, X) \propto p(X|G)p(e_{ij}|e_{-ij}). \tag{3.11}$$

Assuming that under the prior $p(e_{ij}|e_{-ij}) = p(e_{ij})$, sampling from (3.11) is equivalent to sampling from a Bernoulli random variable with the following posterior odds ratio

$$\frac{p(e_{ij} = 1|e_{-ij}, X)}{p(e_{ij} = 0|e_{-ij}, X)} = \frac{p(X|G)}{p(X|G')}\frac{p(e_{ij} = 1)}{p(e_{ij} = 0)} \tag{3.12}$$

where $G' = (V, E')$ with $E' = E \setminus (i, j)$.

Given the nature in which Gibbs moves are proposed, extra non-significant edges tend to be included at each step. See this as follows. Consider that at some point in the chain the current model has all t "important" edges in and all remaining $(T - t)$ non-relevant edges are out. If all remaining edges have the same prior inclusion probability, at each iteration, approximately $(T - t) \times p(e_{ij} = 1)$ are included even when this is a move to models away from regions of high probability. Due to this behavior, Gibbs has a disposition to wonder around areas of low probability, so reducing its ability to efficiently explore high dimensional model spaces. This is a generic problem, exhibited here in the special case of graphical models.

The Metropolis-Hastings algorithm generalize the Gibbs sampler above in the sense that it provides a framework where more flexible specification of moves are allowed. In general, starting from a current graph $G = (V, E)$ a candidate $G' = (V, E')$ is sampled from a proposal distribution $H(G'; G)$ and accepted with probability

$$\alpha = min \left\{ \frac{p(G'|X)}{p(G|X)} \frac{H(G; G')}{H(G'; G)}, 1 \right\}.$$

If an edge e_{ij} is chosen at random and $e'_{ij} = 1 - e_{ij}$ is proposed, the proposal distribution is symmetric, i.e. $H(G'; G) = H(G; G')$ and the acceptance ratio reduces to

$$\alpha = min \left\{ \frac{p(X|G')p(G)}{p(X|G)p(G)}, 1 \right\}. \tag{3.13}$$

This approach was first proposed in a different model selection context by Madigan and York (1995) (MC^3) and used in variable selection and model averaging by Raftery *et al.* (1997) and Clyde *et al.* (1996). It is interesting to point out that this particular Metropolis-Hastings is very similar to the component-wise Gibbs sampler but, as shown by George and McCulloch (1997), the MH is more likely to move at each step, so enhancing its ability to traverse the model space.

In both methods, moves in model space are decided as a function of the ratio of marginal likelihoods between graphs that differ by one edge. If the search is constrained to decomposable models, these ratios can be efficiently computed by the use of (3.10), as described in the previous section. If an unrestricted search is performed, computations cannot be facilitated and the evaluation of the marginal likelihood for a non-decomposable graph has to be approximated by the method described in Section 3.1.1; this induces a huge computational burden in more than trivial dimensions. Constraining the search to decomposable models requires, at each step, the proposed model to be decomposable. In the Gibbs sampler this can be done efficiently (Giudici and Green, 1999); however, in the Metropolis-Hastings this restriction modifies the symmetry property of the proposal distribution, generating the need for additional computations to determine the number of decomposable graphs with one more edge relative to the number of decomposable graphs with one fewer edge. Ignoring this fact creates an irreversible Markov chain and convergence can no longer be guaranteed.

The above MCMC algorithms are closely related and in fact have very similar performance in our experiments. The problem with these methods is that when $p = |V|$ grows, the chain is unable to move much in model space and high probability models tend not to be visited. To try to go beyond this problem I now present a stochastic search method that, due to its parallel nature and design, promotes a more extensive exploration of the model space.

3.5 Shotgun Stochastic Search Algorithm

If Markov chain Monte Carlo is viewed merely as a tool for visiting high probability graphs, there are certainly competing algorithms. The *Shotgun Stochastic Search* (SSS) for Gaussian graphical models is a MCMC inspired approach that takes advantage of distributed computing environment to parallelize computations and expand its ability to search for models. Generally, SSS is a sequential, local move strategy that, at each step, explores the neighborhood of the current model. The algorithm can be summarized as:

(i) Start with a graph G.

(ii) Propose, at random, $\mathcal{F}_1$ neighbors and evaluate the posterior probability (up to a normalizing constant) of each model in parallel. Retain the top $\mathcal{F}_2 \leq \mathcal{F}_1$.

(iii) Sample a new graph from the $\mathcal{F}_2$ top neighbors, with each G_i having probability proportional to p_i^{α}, where p_i is the posterior probability of graph i raised to an annealing parameter $\alpha > 0$.

(iv) Goto step (iii) and iterate.

Most of the computational burden of the above strategy is in step (ii) which can be parallelized with the evaluation of each proposed model probability being carried out independently.

The main idea here is that at every iteration a large number of candidate models are generated, "shooting out" (proposing) moves in various directions. This method can be thought as a Metropolis-Hastings that incorporates information from the likelihood into the proposal resulting in a chain that can move faster to areas of high probability. A detail discussion of the connections between SSS and MCMC in the context of regression models appears in Hans (2005).

To keep the discussion consistent, in experimenting with SSS, "neighbors" were defined as graphs differing by only one edge and at each step all neighbors were evaluated with $\mathcal{F}_1 = \mathcal{F}_2 = p(p-1)/2$. If the search is restricted to decomposable graphs, the only necessary modification is to constraint the proposed neighborhood to decomposable models which can be efficiently implemented (see Appendix A, and Giudici and Green, 1999). In this case, the evaluation of posterior probabilities is facilitated by local updates of marginal likelihoods ratios as described in section 3.3.

3.6 Examples

In order to understand the performance and scalability of MCMC and SSS methods two moderate size simulated examples are presented, along with a large scale 150 node example from gene expression studies. In each case, restricted (decomposable models only) and unrestricted searches were performed. Given the similar performance of the two MCMC methods, the results presented here focus on the Metropolis-Hastings (MH) algorithm. Throughout the examples, hyper-inverse Wishart priors were used with parameters b set to very small values and $D = \tau I$. This choice is somehow non-informative and consistent with problems in which variables are measure in the same scale of variation so, for simplicity, all datasets are standardized.

First, consider two simulated examples where the underlying graph is known. Figure 3.4 shows a 15 node decomposable graph whereas Figure 3.5 displays a 12 node non-decomposable graph in a single non-complete prime component. Each data set consists of 250 observations so there are no constraints on the maximum prime component size. The first simulated dataset was inspired by patterns of daily currency exchange fluctuations against the US dollar. Consequently, the data ranges approximately between $\pm 2\%$. This range is about two standard deviations, so $\sigma_{ii}^2 \approx 0.0001$ hence the choice of $b = 3$ and $\tau = 0.0001$ so the prior expected value for the variance σ_{ii} is approximately equal to the data variance. For the second dataset, $\mathbf{\Sigma}$ is actually a random draw from the $HIW_G(3, I)$ so $\tau = 1$ and $b = 3$. In both cases the prior over graphs is the sparsity encouraging prior suggested before with $\beta = 2/(|V| - 1)$. The annealing parameter α in the SSS was set at 1.0 and had no real impact in these small examples.

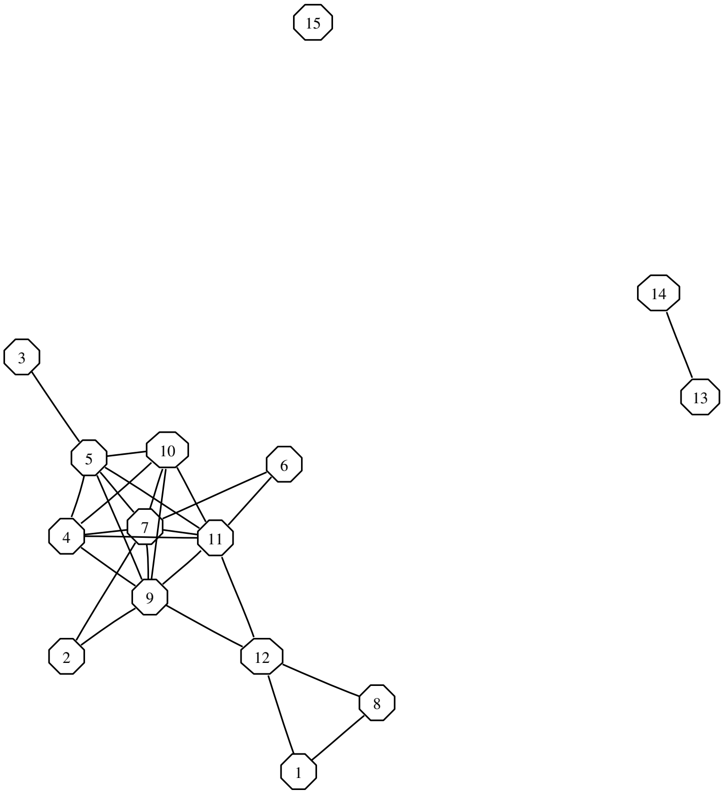


Figure 3.4: The true underlying decomposable graph on $p = 15$ nodes – the first simulated example

For each example, the Metropolis-Hastings was run for 10,000$\times\binom{|V|}{2}$ steps (where $\binom{|V|}{2}$ is the number of possible one edge moves in the unrestricted case). The shotgun stochastic search algorithm was run 10,000 iterations; at each iteration all possible (unrestricted) one edge moves were considered, so to perform the same number of graph comparisons as the Metropolis-Hastings algorithm. In both scenarios the empty graph (no edges) was used as a starting point.

The algorithms clearly use a similar amount of computing resources, as they evaluate the same number of comparisons between current and proposed graphs. However, the stochastic search algorithm is parallelizable. The run times for both types of algorithm are given in Tables 3.1 and 3.2. The Metropolis-Hastings was run on a Dell PC with a 1.8 MHz Xeon processor in a Linux environment, and the shotgun stochastic search on a Beowulf cluster with 26 dual processor, 1.4Mhz nodes. The C++ implementation for all methods used are available at www.isds.duke.edu under the *Software* link.

In both examples the MH and SSS identify the same top model in a short amount of time. The top graph is defined as the graph with highest posterior probability of all graphs visited. Note that in this small example the MH runs faster than SSS. This is an artifact of this small example given by the computational overhead of SSS (initialization of many processors). When the problem grows in complexity or dimension, SSS becomes much more efficient. This is also seen in the results of the unrestricted search.

The top decomposable graphs – those identified with highest posterior probability – are displayed in Figures 3.6 and 3.7; the top graphs from the unrestricted

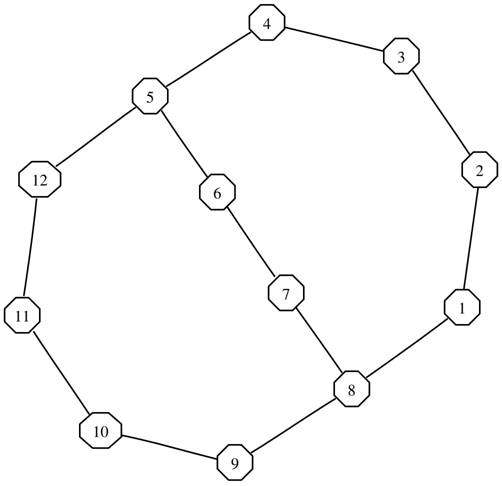


Figure 3.5: The true underlying non-decomposable graph on $p = 12$ nodes – the second simulated example

Method	Runtime (secs)	Max log posterior	Graphs to first top graph visit	Time to first top graph visit
MH-d	36	-2591.18	912	1
SSS-d	183	-2591.18	792	2
MH-u	15,220	-2590.94	415	2
SSS-u	2773	-2590.94	13,266	5

Table 3.1: Comparison between algorithms runtime, and quality of best graph found, for the 12 node example. MH-d(u) refers to the Metropolis-Hastings algorithm on decomposable (unrestricted) models, while SSS-d(u) refers to the shotgun stochastic search method on decomposable (unrestricted) models.

Method	Runtime (secs)	Max log posterior	Graphs to first top graph visit	Time to first top graph visit
MH-d	93	15633.76	349,484	36
SSS-d	234	15633.76	33,495	9
MH-u	513,077	15633.83	666,425	309,222
SSS-u	5930	15636.38	82,845	112

Table 3.2: Comparison between algorithms runtime, and quality of best graph found, for the 15 node example. MH-d(u) refers to the Metropolis-Hastings algorithm on decomposable (unrestricted) models, while SSS-d(u) refers to the shotgun stochastic search method on decomposable (unrestricted) models.

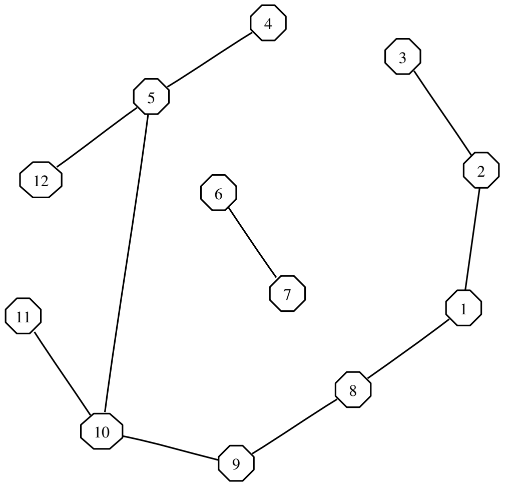


Figure 3.6: Highest log posterior graph for the 12 node example when the search is restricted to decomposable models.

search appear in Figures 3.8 and 3.9. Likelihood comparison with true graphs show that each of these graphs have greater likelihood as well as posterior probability than the true graph. Also, in both examples, the most probable graph found was insensitive to the starting point. The same graphs were found starting at the full graph.

A more challenging problem is the analysis of DNA microarray derived gene expression data from $p = 150$ genes in $n = 49$ samples, associated with the estrogen receptor pathway, from the study of West *et al.* (2001). The data was standardized and the prior specified with $b = 3$ and $\tau = 4$. Due to the small number of observations relative to the number of variables, there is the additional constraint that the maximum prime component size cannot exceed $n - 1$. Finally, in this context, the sparsity encouraging prior can be interpreted as a belief that on average, each gene has major interactions with a relatively small number of other genes. The results from three algorithms are shown in Table 3.3. Times are now given in hours. The unrestricted search Metropolis-Hastings had a very poor performance so their results are omitted here. The best results for the shotgun search algorithm were obtained with an annealing parameter of 50, which essentially represents a deterministic hill climbing in the space of graphs (see Figure 3.10). In this large example the advantage of SSS is enormous as it is able to get to areas of higher probability in the model space much sooner than any MCMC method.

In the unrestricted case, increasing the number of Monte Carlo samples in order to get a sharp enough evaluation of the marginal likelihood was not feasible; settling for a standard deviation of the log likelihood of 1.0 resulted in one cycle

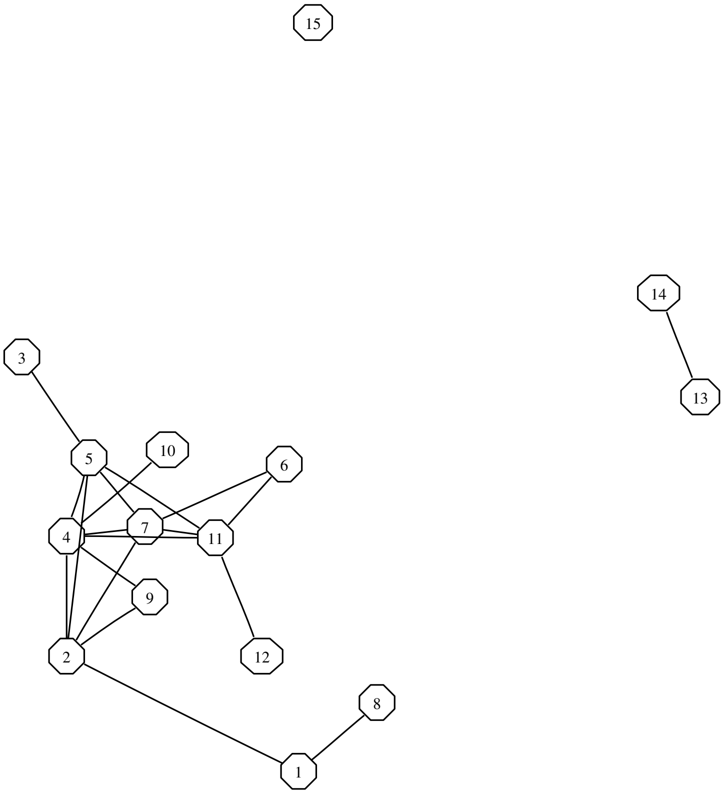


Figure 3.7: Highest log posterior graph for the 15 node example when the search is restricted to decomposable models.

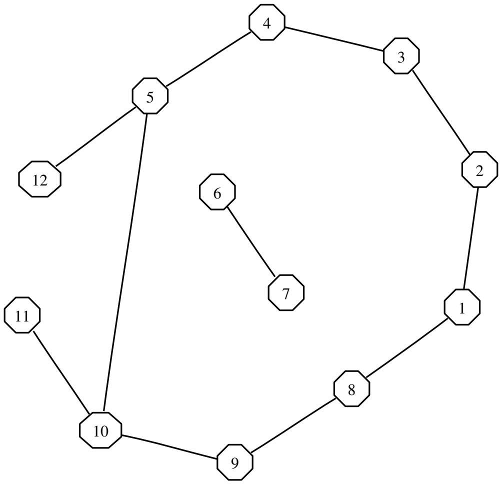


Figure 3.8: Highest log posterior graph for the 12 node example when the search is unrestricted

of neighbors evaluations (a single step in our stochastic search procedure) taking up to 40 computer days (1 day on a 40 node cluster). Using this procedure, starting from the empty graph and running until the estimated log posterior stopped improving, the best graph found had log posterior −9364.67, worse than the best decomposable graph found in the restricted search. This graph may represent a local mode not present in the decomposable framework, or be the result of suboptimal moves resulting from the imprecise Monte Carlo approximation of the marginal likelihood. Table 3.3 shows the best graph found by starting at the best decomposable graph (the final estimate of the log posterior for this graph was run with enough Monte Carlo samples to put the MC standard deviation below 0.1). A total of 10 cycles of evaluating all neighbors were done. As these graphs were "close" to decomposable graphs, the evaluation time was reduced versus graphs with similar numbers of edges produced by the search starting at the empty graph.

3.6.1 Difficulties evaluating Non-Decomposable Models

As seen in Figure 3.1 and also indicated by the example above, estimating the marginal likelihood for large scale non-decomposable graphs is a very demanding task. High variance of the Monte Carlo based marginal likelihood estimator promotes "artificial" moves and interferes with the ability to explore the model space.

To understand the behavior of the normalizing constant estimates, a simulation study on non-complete prime components with different numbers of nodes is performed. Two examples for each size were selected from those that occurred

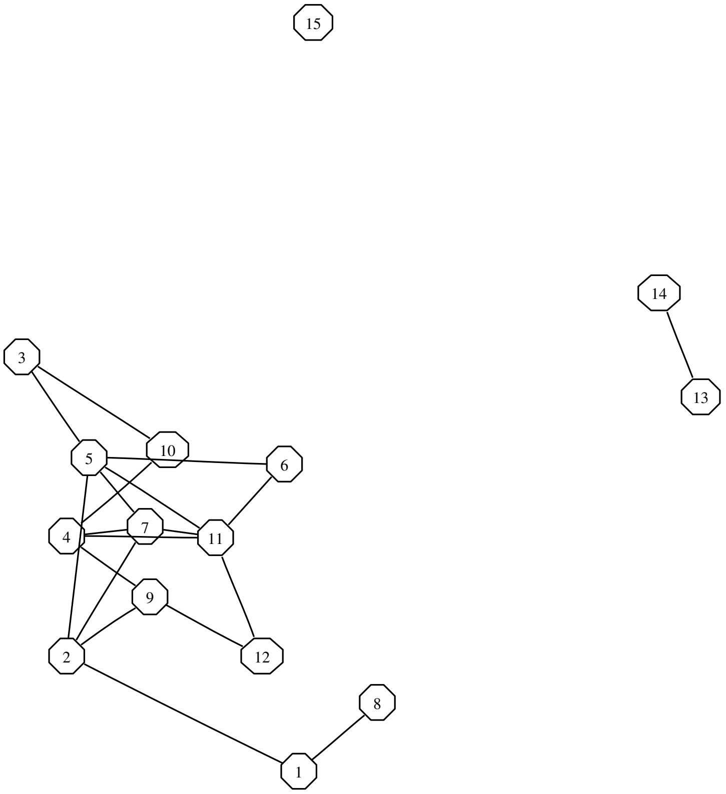


Figure 3.9: Highest log posterior graph for the 15 node example when the search is unrestricted.

during the MH model search for the 15 variable data set. Given that the search strategies depend on likelihood ratios, it is the variance of the log normalizing constants that are relevant here. Figures 3.11(A) and 3.11(B) show the variances of the estimated log of the prior and posterior normalizing constants (where the estimate is based on 100 random draws). The plotted variances are of course estimates themselves, each based on 1000 separate normalizing constant estimations.

The estimate of the log prior normalizing constants have systematically smaller variances than the corresponding estimates for the posterior; there is also a tendency for variance to increase with component size. This can be partially explained by examining the form of $\boldsymbol{\psi}$, the sampled matrix from which the estimate is computed. The variance of diagonal entries ψ_{ii} increase as one moves down the diagonal, so in larger prime components, more uncertainty is associated with $\boldsymbol{\psi}$.

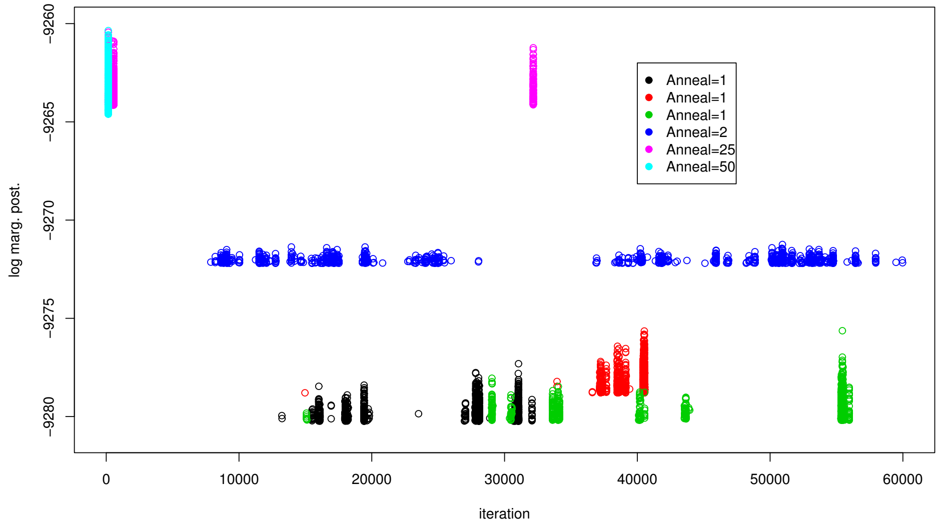


Figure 3.10: Exploration of the model space with different annealing parameters. The figure plots the posterior probability for the top 1000 graphs visited in each search against the iteration where the graph was visited.

Method	Runtime (hrs)	Max log posterior	Graph to first top graph visit	Time to first top graph visit
MH-d	18.02	-9417.97	100,466,818	6.51
SSS-d	0.03	-9260.84	1,698,600	0.03
SSS-u	6.29*	-9227.68	44,700	3.39

Table 3.3: Comparison between algorithms runtime, and quality of best graph found, for the gene expression example. *Starting from the best decomposable graph found. MH-d refers to the Metropolis-Hastings algorithm on decomposable models, while SSS-d(u) refers to the shotgun stochastic search method on decomposable (unrestricted) models.

It also possible to note that the ordering of the variables used when setting up $\boldsymbol{\psi}$ affects the variance of the log normalizing constants. Each prime component considered in Figure 3.11(C) is a cycle; in the "optimal" configuration, each variable, except the first and the last, has exactly one neighbor preceding it in the rows of $\boldsymbol{\psi}$. The "worst" configuration has the first $|V|/2$ variables each with both neighbors occurring further down in the matrix.

The cause of this phenomenon can be seen by factoring Equation (3.6) into the constant C_b, and the part estimated by Monte Carlo, $\mathbf{E}(f(\boldsymbol{\psi}))$. Recall that t_{ii} are the entries of the Cholesky decomposition of the HIW parameter $\boldsymbol{\psi}$ ($\boldsymbol{\psi}^*$ for the posterior), ν_i is the number of neighbors of node i subsequent to it in the ordering of vertices, and z_i is the total number of neighbors of node i, plus one. We list the variables of a prime component in an arbitrary order, but the relative sizes of C_b and $\mathbf{E}(f(\boldsymbol{\psi}))$ clearly depend on the ordering of the variables (although their product is constant–the expression is valid for any ordering). In the experiments the variance of $\bar{f}(\boldsymbol{\psi})$ was (roughly) unaffected by ordering; however, the interesting quantity is the variance of $\log(\bar{f}(\boldsymbol{\psi}))$ which did suffer from ordering effects. The "optimal" ordering for cycles discussed above minimizes C_b for the HIW prior; the "worst" ordering maximizes it. Although the order of variables in a prime component influence the estimation of its normalizing constant, trying to optimize the ordering at each step is not an attractive alternative due to the computational cost involved with it.

The highest variance samples in Figure 3.11(B) represent very low likelihood graphs, which have small $\mathbf{E}(f(\boldsymbol{\phi}))$– and high variance $\log(\mathbf{E}(f(\boldsymbol{\phi})))$–regardless of ordering. Figure 3.11(D), a plot of variances of log posterior normalizing constants for prime components in graphs *accepted* during the MH search, is more consistent with the variance trends in the log prior normalizing constants. The variance of the "worst" case for each component size seems to be a function of the size of the component considered, $|V_P|$. Based on this, $1.5|V_P|^3$ samples for the posterior normalizing constants and $0.5|V_P|^3$ for the prior normalizing constants were used. This scheme solved the problem with the chain mobility discussed at the beginning of this section. To be safe, at the end of each unrestricted model search, all graphs with a log posterior within 2.0 of the top log posterior were reexamined with enough Monte Carlo samples to ensure the graph listed as "best" did indeed have the highest log posterior.

3.7 Discussion

After experimenting much further with the methodology described in this chapter, it became clear that exploring the model space of decomposable Gaussian graphical models with a large number of variables, certainly up to a few hundreds, is feasible. As I hope to have illustrated with some of the examples, traditional MCMC methods are only competitive in relatively small problems whereas stochastic search methods, such as SSS, are definitely successful in identifying high probability areas in larger model spaces. Local-move type methods are really advantageous for decomposable graphs where not only are computations exact, but they can also be efficiently implemented given some of the properties of such models.

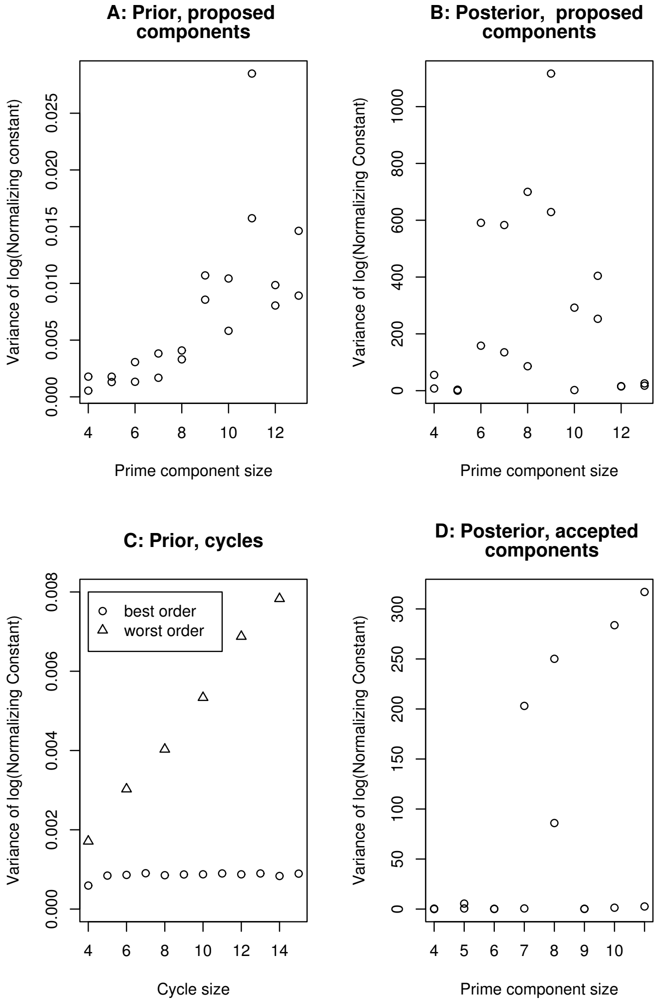


Figure 3.11: Relationship between the variance of the estimated normalizing constants, based on 100 samples, and the size of the prime component. Four cases are considered: (A), the prior normalizing constant for components proposed during the unrestricted model search for the 15 node dataset, (B), the posterior normalizing constants for these components, (C), prior normalizing constants for cycles, using different variable orderings, and (D), posterior normalizing constants for components considered during the unrestricted model search and subsequently accepted by the Metropolis-Hastings algorithm.

When all graphs are considered, life is more complicated. Experience shows that in small examples, up to 20 variables, model search can be accomplished but it becomes very challenging quickly thereafter. The necessity to estimate marginal likelihoods for possibly large prime components create a huge computational burden that undermine our ability to search for models and other alternatives are needed to address this problem. Local search of unrestricted graphs around "good" decomposable graphs or other candidate graphs generated interesting results in the 150 node example and represents a promising strategy.

The main contribution of this chapter has been to present and thoroughly experiment with a shotgun stochastic search algorithm designed for explore the very high-dimensional discrete model space to identify regions of high probability and the corresponding graphs. The approach is parallelizable and this work also serves to explore and evaluate distributed implementation. More experimentation with the annealing schedules is needed to find optimal strategies for different situations. In the 150 node example deterministic hill climbing produced the best results in terms of rapid identification of high probability graphs. Turning the temperature "up" and "down" at different times could promote a better exploration of the model space.

It is apparent that significant near-term progress in model and variable selection in the face of higher-dimensional problems is unlikely if computations are restricted to single processors. The hope of our experiments is to show that distributed computation are essential to the development of more efficient search methods.

Chapter 4

Simulation of the Hyper-Inverse Wishart Distribution

A central element in the analysis of Gaussian graphical models is the hyper-inverse Wishart (HIW) distribution. For computational reasons, sampling from this class of distributions has been avoided in prior work. For example, in the RJMCMC analysis of Giudici and Green (1999), where the covariance matrix has to be updated in every step, importance sampling methods are used. Roverato (2000) suggests an alternative parametrization, based on the Cholesky decomposition of the precision matrix, that would provide a way to sample from HIW distribution in decomposable graphs. In large-scale problems his method requires the Cholesky decomposition of large matrices and so rapidly becomes unattractive. In application of Gaussian graphical models we often require inference for complicated functions of the parameters of a variance matrix and so an approach to direct simulation of posteriors under HIW models is of serious practical value.

In this chapter I define and exemplify an efficient method for direct simulation of structured random matrices under an HIW distribution for both decomposable and non-decomposable models. The strategy uses the junction tree of a graph to decompose the HIW distribution and thus allows us to work sequentially at the prime component level. In decomposable models this decomposition provide access to standard distributional theory for the inverse Wishart distribution. In the non-decomposable case, standard distributional results no longer hold and properties of the inverse of HIW distributions (Atay-Kayis and Massam, 2005) are used. The ability to sample directly from the hyper-inverse Wishart allows direct Monte Carlo computations for detailed posterior inference on elements of covariance matrices and functions of them; I illustrate this in an example drawn from portfolio analysis in financial time series.

4.1 Simulation Method

The sampling strategy is based on the compositional form of the joint distribution over the sequence of subgraphs defined by the junction tree.

Let $G = (V, E)$ be a graph on p nodes and assume a Gaussian graphical model with $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} \in M(G)$. Suppose that $\mathbf{\Sigma} \sim HIW_G(b, \mathbf{D})$. By generating the junction tree of G, prime components are perfectly ordered as $\{P_1, S_2, P_2, \ldots, P_k\}$, and the joint density factorizes as

$$p(\mathbf{\Sigma}|b, \mathbf{D}) = p(\mathbf{\Sigma}_{P_1}) \prod_{i=2}^{k} p(\mathbf{\Sigma}_{P_i}|\mathbf{\Sigma}_{S_i}). \tag{4.1}$$

Equation (4.1) is a direct consequence of property (2.1) and indicates that, starting from $\mathbf{\Sigma}_{P_1}$, there is a clear sequence of conditional distributions to be simulated in order to obtain a draw from $p(\mathbf{\Sigma}|b, \mathbf{D})$ via composition. We simply need to identify the sequence of conditional distributions and a method to sample them.

4.1.1 Decomposable Models

In decomposable models where all prime components are complete, conditioning results for inverse-Wishart random variables enable sampling from each of the elements in the composition directly.

For a perfect ordering of cliques and separators $\{P_1 = C_1, S_2, P_2 = C_2, \ldots, S_k, P_k = C_k\}$ we use the traditional notation $R_i = C_i \setminus S_i$ and write $\mathbf{\Sigma}_{C_i}$ and $\mathbf{D}_{C_i}$ in their conformably partitioned forms

$$\mathbf{\Sigma}_{C_i} = \begin{pmatrix} \mathbf{\Sigma}_{S_i} & \mathbf{\Sigma}_{S_i,R_i} \\ \mathbf{\Sigma}_{R_i,S_i} & \mathbf{\Sigma}_{R_i} \end{pmatrix} \quad \text{and} \quad \mathbf{D}_{C_i} = \begin{pmatrix} \mathbf{D}_{S_i} & \mathbf{D}_{S_i,R_i} \\ \mathbf{D}_{R_i,S_i} & \mathbf{D}_{R_i} \end{pmatrix}$$

where $\mathbf{\Sigma}_{S_i,R_i} = \mathbf{\Sigma}_{R_i,S_i}^T$. Also, let

$$\begin{aligned} \mathbf{\Sigma}_{R_i.S_i} &= \mathbf{\Sigma}_{R_i} - \mathbf{\Sigma}_{R_i,S_i}\mathbf{\Sigma}_{S_i}^{-1}\mathbf{\Sigma}_{S_i,R_i} \\ \mathbf{D}_{R_i.S_i} &= \mathbf{D}_{R_i} - \mathbf{D}_{R_i,S_i}\mathbf{D}_{S_i}^{-1}\mathbf{D}_{S_i,R_i}. \end{aligned}$$

The sampling scheme is defined as:

(i) Sample $\mathbf{\Sigma}_{C_1} \sim IW(b, \mathbf{D}_{C_1})$; this also gives values to the submatrix $\mathbf{\Sigma}_{S_2}$.

(ii) For $i = 2, \ldots, k$, sample

$$\begin{aligned} \mathbf{\Sigma}_{R_i.S_i} &\sim IW(b + |R_i|, \mathbf{D}_{R_i.S_i}) \qquad \text{and} \\ \mathbf{U}_i &\sim N(\mathbf{D}_{R_i,S_i}\mathbf{D}_{S_i}^{-1}, \mathbf{\Sigma}_{R_i.S_i} \otimes \mathbf{D}_{S_i}^{-1}). \end{aligned}$$

Then directly compute the implied values of $\mathbf{\Sigma}_{R_i,S_i} = \mathbf{U}_i\mathbf{\Sigma}_{S_i}$ and $\mathbf{\Sigma}_{R_i} = \mathbf{\Sigma}_{R_i.S_i} + \mathbf{\Sigma}_{R_i,S_i}\mathbf{\Sigma}_{S_i}^{-1}\mathbf{\Sigma}_{S_i,R_i}$.

This sequence completes the sampling of all elements in the intersecting block components of $\mathbf{\Sigma}$ on the junction tree. It remains to fill-in the implied values of the elements of $\mathbf{\Sigma}$ in the positions where $\omega_{ij} = 0$, i.e. $(i, j) \notin E$. This is done via the standard completion operation described in (2.11).

4.1.2 Non-Decomposable Models

In non-decomposable models the same junction tree representation for compositional sampling is used, so breaking the problem into a series of conditional simulations. The steps are precisely as described above for prime components that are complete. The key difference and computational difficulties arise when a non-complete prime component is visited; for such a component the standard conditioning results for the inverse-Wishart (step (ii) above) do not apply. The challenge is then to identify a way to sample from the appropriate conditional distribution of the elements of $\mathbf{\Sigma}$ in that component conditional on the set of values on its preceding separator.

Here I use and extend the general theory of Atay-Kayis and Massam (2005) that expresses a (global) HIW distribution through the Cholesky decomposition of $\mathbf{\Omega}$. The key points here are: *(a)* to use this only in each incomplete prime component within the overall compositional sampler, so allowing for efficient computation and scaling to large graphical models by exploiting local computation; and *(b)* to extend the theory to derive samples from the *conditional* distributions of HIW matrices given separating parameters. The details are as follows.

For any incomplete prime component P, first consider the Cholesky method for sampling a defined HIW distribution $\mathbf{\Sigma}_P \sim HIW_P(b, \mathbf{D}_P)$ on that component alone. Write $\mathbf{D}_P^{-1} = \mathbf{T}^t\mathbf{T}$ for the Cholesky decomposition of the HIW parameter matrix. Then, for $\mathbf{\Omega}_P = \mathbf{\Sigma}_P^{-1}$ write the Cholesky decomposition as $\mathbf{\Omega}_P = \mathbf{\Phi}^t\mathbf{\Phi}$, and define $\mathbf{\Psi} = \mathbf{\Phi}\mathbf{T}^{-1}$. The structure of the subgraph P implies certain constraints on the elements of $\mathbf{\Psi}$ (Atay-Kayis and Massam, 2005; Jones *et al.*, 2005); the *free* elements are the ψ_{ii} and those ψ_{ij} such that (i, j) is an edge in P, and can be simulated directly from independent chi-square and normal random variates.

Then $\mathbf{\Psi}$ will be completed by direct, deterministic evaluation of the remaining constrained elements. In detail:

(i) Compute the Cholesky decomposition $\mathbf{T}$ of $\mathbf{D}_P^{-1}$.

(ii) Define $t_{\langle ij]} = t_{ij}/t_{jj}$.

(iii) Create the $p \times p$ upper triangular matrix $\mathbf{A}$ with $a_{ii} = 0$ and, for $i \neq j$, $a_{ij} = 1$ if (i, j) is an edge in P, zero otherwise.

(iv) Compute ν_i as the number of 1's in the i^{th} row of $\mathbf{A}$.

(v) Sample the free variables ψ_{ij} for edges (i, j) in P:

- for $i = 1$ to p, $\psi_{ii} = \sqrt{u_i}$, where $u_i \sim \chi^2_{b+\nu_i}$,
- for $i \neq j$ and $a_{ij} = 1$, $\psi_{ij} \sim N(0, 1)$.

For edges (i, j) not in P, compute ψ_{ij} via:

- if $i = 1$,

$$\psi_{ij} = -\sum_{k=1}^{j-1} \psi_{ik} t_{\langle kj]},$$

- and for $i > 1$,

$$\psi_{ij} = \sum_{k=i}^{j-1} \psi_{ik} t_{\langle kj]} - \sum_{r=1}^{i-1} \left(\frac{\psi_{ri} + \sum_{l=r}^{i-1} \psi_{rl} t_{\langle li]}}{\psi_{ii}} \right) \left(\psi_{rj} + \sum_{l=r}^{j-1} \psi_{rl} t_{\langle kj]} \right).$$

(vi) Finally, set $\mathbf{\Phi} = \mathbf{\Psi}\mathbf{T}$ and compute $\mathbf{\Omega}_P = \mathbf{\Phi}^t\mathbf{\Phi}$, hence delivering $\mathbf{\Sigma}_P = \mathbf{\Omega}_P^{-1}$.

Now, the modification needed is that we want to sample from $p(\mathbf{\Sigma}_P|\mathbf{\Sigma}_S)$ where S represents the nodes in P that lie in the preceding separator in the junction tree, so that $\mathbf{\Sigma}_S$ is an upper left block of $\mathbf{\Sigma}_P$ as in Section 4.1.1. In order to define the modification the following two Lemmas are useful:

Lemma 4.1. *If $\mathbf{\Omega} \in M(G)$ follows a HIW_G and given a perfect order of prime components $\{P_1, S_2, \ldots, P_k\}$ the matrix $\mathbf{\Phi}$ defined as $\mathbf{\Omega} = \mathbf{\Phi}^t\mathbf{\Phi}$ can be compactly specified as $\{\mathbf{\Phi}_{P_1}, (\mathbf{\Phi}_{R_2}\mathbf{\Phi}_{S_2R_2}), \ldots, (\mathbf{\Phi}_{R_k}\mathbf{\Phi}_{S_kR_k})\}$ where all elements in the sequence are mutually independent. This implies that, for all $i = 1, \ldots k$,*

$$\mathbf{\Phi}_{P_i} \perp\!\!\!\perp \mathbf{\Phi}_{P_{i+1}} | \mathbf{\Phi}_{S_{i+1}}.$$

Proof. See Roverato (2000) □

Lemma 4.2. *Let $\mathbf{D} \in M(G)$. Given a perfect order $\{P_1, S_2, \ldots, P_k\}$ we can write*

$$\mathbf{D} = \mathbf{L}^t\mathbf{L} = \sum_{P \in \mathcal{P}} \mathbf{D}_P - \sum_{S \in \mathcal{S}} \mathbf{D}_S$$

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

$$\mathbf{D}_{P_j} = \left(\begin{array}{cc} \mathbf{L}_{S_j} & 0 \\ \mathbf{L}_{R_jS_j} & \mathbf{L}_{R_j} \end{array} \right) \left(\begin{array}{cc} \mathbf{L}^t_{S_j} & \mathbf{L}_{S_jR_j} \\ 0 & \mathbf{L}^t_{R_j} \end{array} \right)$$

Proof. This is a direct consequence of the collapsibility of graph G with graphs G_{H_j} (Lauritzen, 1996; Roverato, 2000). □

Due to Lemmas 4.1 and 4.2 the necessary changes are in fact almost trivial: from Lemma 4.1 it is possible to note that the remaining elements of $\mathbf{\Phi}$ in the current prime component are independent of every element in H given the elements in S; Therefore, conditioning is equivalent to fixing the values of the elements in the initial rows of $\mathbf{\Phi}$ (and therefore $\mathbf{\Psi}$) corresponding to the separator S, and skipping the corresponding steps in the sequence of computations above. Lemma 4.2 allow

1	New Zealand Dollar (NZD)
2	Australian Dollar (AUS)
3	Japanese Yen (JPY)
4	Swedish Krone (SEK)
5	British Pound (GBP)
6	Spanish Peseta (ESP)
7	Belgian Franc (BEF)
8	French Franc (FRF)
9	Swiss Franc (CHF)
10	Dutch Guilder (NLG)
11	German Mark (DEM)

Table 4.1: Eleven international currencies. The dataset consists of 2566 daily returns for each of these 11 currencies, over the period of about 10 years – 10/9/86 to 8/9/96.

us to maintain local computations, as the elements of $\mathbf{\Phi}$ corresponding to S can be obtained from the Cholesky of the $\mathbf{\Sigma}$ elements in the preceding prime component alone; then the corresponding elements of $\mathbf{\Psi}$ can be immediately computed and plugged-in step (v) above.

After sampling $\mathbf{\Sigma}_P$, we continue moving through the junction tree, working with both complete (cliques) or incomplete prime components until all the block components of the full $\mathbf{\Sigma}$ are completed. Then, again as described in Equation (2.11), the completion operation comes in to play to fill-in the remaining elements of $\mathbf{\Sigma}$.

4.1.3 Examples

An example concerns posterior inference on a 11-dimensional covariance matrix based on the graph G (Figure 4.1) linking international currency exchange rates relative to the US dollar (Table 1). Data on consecutive daily returns provide $n = 100$ observations that, after mean-centering and scaling, provide a sum of squares matrix **S**. The graph is immediately interpretable from an econometric finance viewpoint, and consistent with prior data. The graph also happens to be

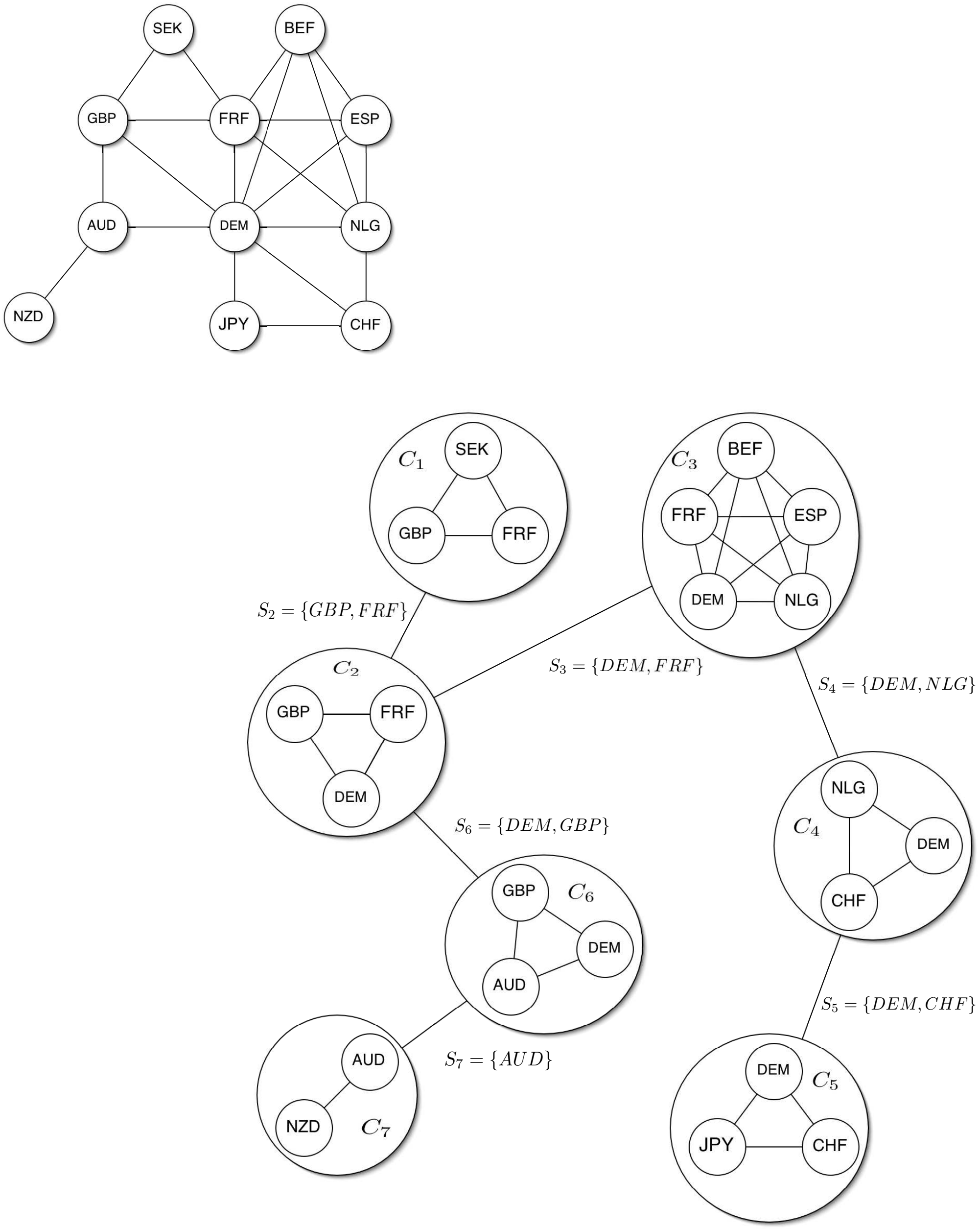


Figure 4.1: Graph and junction tree in exchange rate/portfolio investment example.

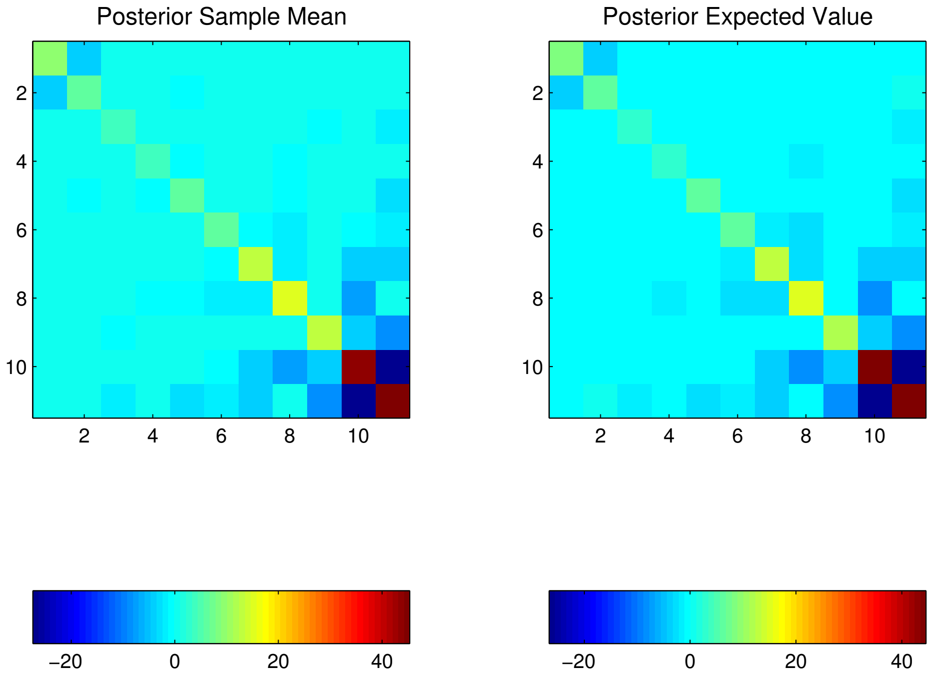


Figure 4.2: Images of the MCMC estimate of posterior mean of Ω (left image) and the theoretically exact posterior mean of Ω (right image) in the exchange rate example.

decomposable, so that under a specified HIW prior $\mathbf{\Sigma} \sim HIW_G(b_0, \mathbf{D}_0)$ the implied posterior is the decomposable HIW form $(\mathbf{\Sigma}|n, \mathbf{S}) \sim HIW_G(b, \mathbf{D})$ with $b = b_0 + n$ and $\mathbf{D} = \mathbf{D}_0 + \mathbf{S}$. The prior parameters chosen are relatively non-informative, with $b = 3$ and $\mathbf{D} = 0.0001I$. We make comparisons below with a parallel analysis on the full graph, i.e., under the usual full inverse Wishart distribution with no conditional independence constraints - so ignoring econometric structuring and also the parsimony that is embodied in G. The difference in log-marginal likelihood of G to the full graph is 102.6, so indicating that the current $n = 100$ observations very, very strongly supports the structured graph relative to the full graph (even ignoring numbers of parameters and the issue of parsimony).

The simulation method described was applied to generate 1000 samples from

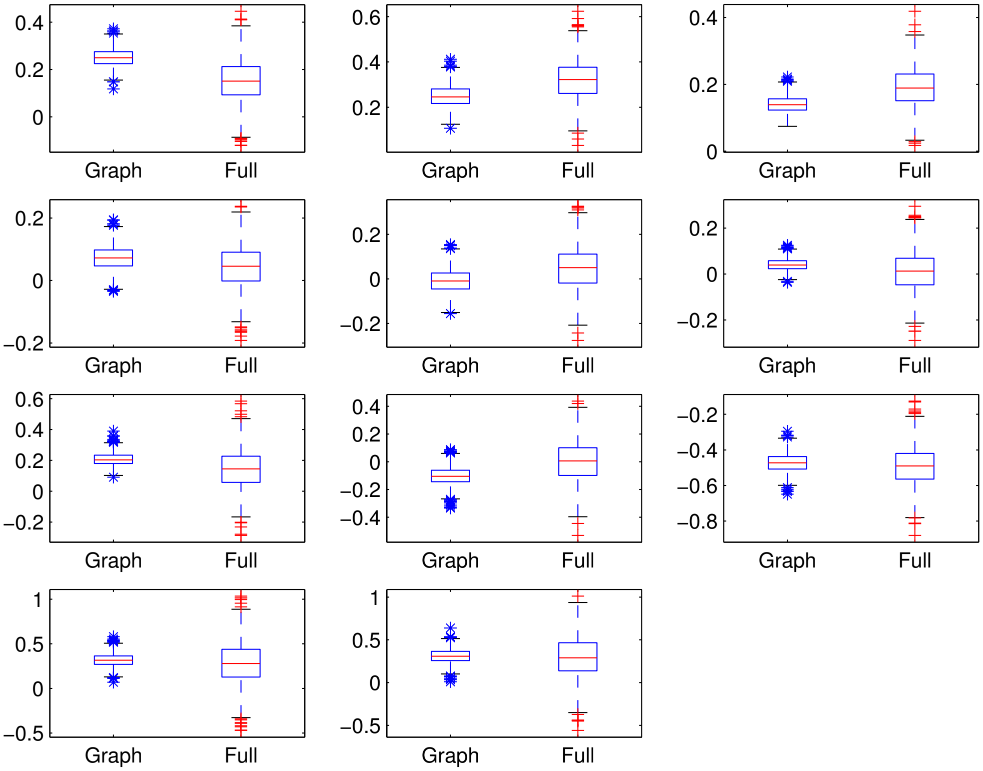


Figure 4.3: Boxplot summaries of posterior distributions of optimal portfolio weights a under G (graph) and the competing full graph (full).

the posterior HIW. Figure 4.2 images the theoretically exact value of $E(\mathbf{\Omega}|n, \mathbf{S})$ and compares it to the image of the Monte Carlo estimate - the latter being just the sample mean of the 1000 simulated precision matrices. The comparison can be investigated in more detail but the graphs suffice to demonstrate the efficacy of the simulation.

Of central practical importance in financial times series and portfolio management are functions of variance matrices of (residual) returns that define optimal portfolio reallocations in sequential decision making about investments such as on exchange rates (Aguilar and West, 2000; Quintana *et al.*, 2003). This serves as a very nice and practically linked example of inference on functions of variance-

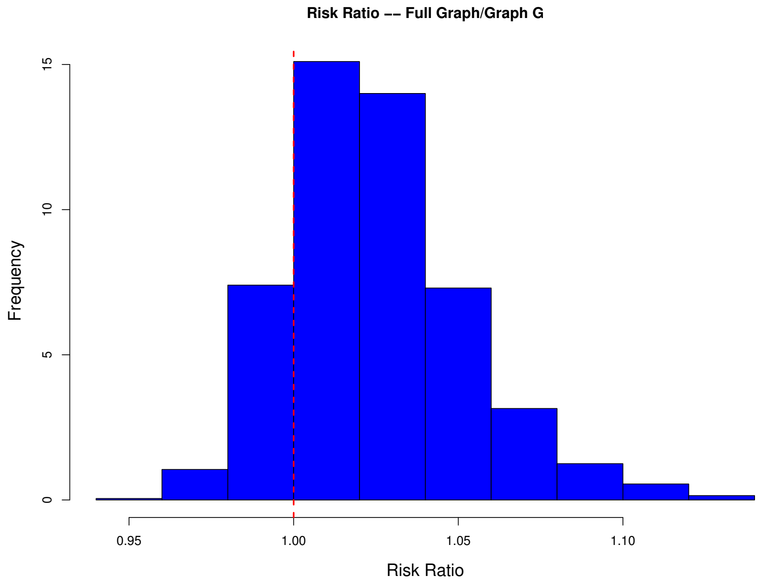


Figure 4.4: Posterior distribution (in terms of histogram of posterior simulated values) of the ratio of standard deviations of the optimal portfolios under the full graph relative to that under the graph G. This indicates that, for a common target return, the risk of the optimal portfolio strategy using the standard, full model is likely to be substantially higher than under the graphical model, which also dominates in terms of fit to the data.

covariance parameters and the use of simulation of structured models of variance matrices. If y represents the returns at the next time point, and a is a vector of 11 weights representing proportional allocation of funds invested in each of the 11 currencies, then the constrained $(1'a = 1)$ portfolio minimizing standard deviation as a measure of risk is given by the choice $a = \mathbf{\Omega}1/(1'\mathbf{\Omega}1)$ (Aguilar and West, 2000). The corresponding risk level is $SD(a'y) = 1/\sqrt{(1'\mathbf{\Omega}1)}$. Hence posterior samples of $\mathbf{\Omega}$ produce, by direct computation, posterior samples for the optimal portfolio weights and related minimized risk.

Figures 4.3 and 4.4 summarize these posterior samples from the HIW posterior on the structured graph G using, as a benchmark comparison, parallel analysis

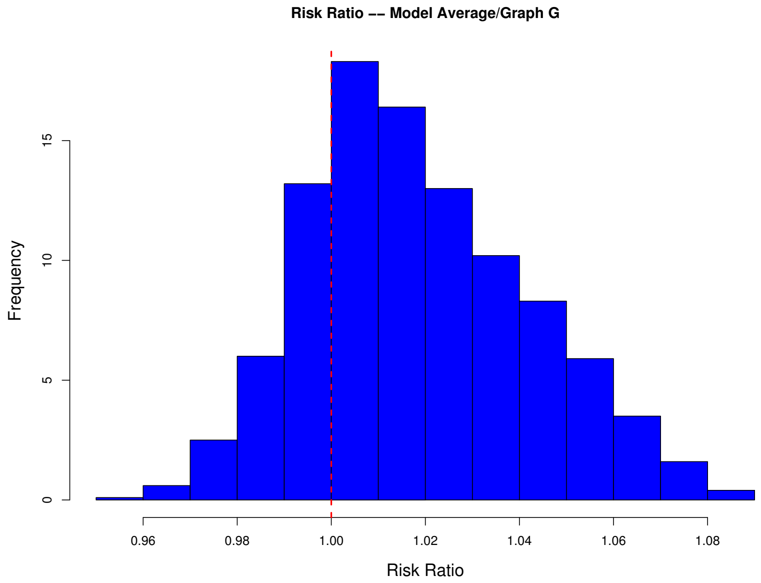


Figure 4.5: Posterior distribution of the ratio of standard deviations of the optimal portfolios under model average relative to that under the graph G.

on the full (unconstrained) graph that would typically be used. Two practically relevant conclusions are apparent from these figures. First, Figure 4.3 shows that the values and levels of variation of the optimal portfolio weights across currencies are smaller than under the full model, implying a more stable investment portfolio of a kind that is desirable on economic and business grounds (Ledoit and Wolf, 2004). Secondly, Figure 4.4 shows that the optimal risk level is inferred as likely to be smaller - and, practically significantly smaller - under the graph G. This forcefully suggests that a structured, parsimonious graphical model can indeed aid in reducing uncertainty and variation in portfolio weights, and thereby reduce investment risk.

To take this further we combine variance matrix parameter learning with learning about the graphical model using results from the MCMC search over graphs too. From that search, the 20 most probable graphs identified appear to have posterior probabilities substantially exceeding that of other discovered graphs (over 100 units in the log-likelihood scale), so that uncertainty about the graph structure may be approximately represented by these 20 graphs; the graph G is the posterior modal graph. Under a formal model averaging strategy, the uncertainty about graphs feeds through to the posterior distribution for the portfolio weights and variance (risk), and these can be compared with the portfolios from both the graph G and the full graph already described. The computations then use the HIW simulator for the posteriors conditional on each of the sampled graphs, and average results with respect to the evaluated posterior probabilities of those graphs. The results appear in Figures 4.5 and 4.6. Evidently, the projected portfolio risk under this "Bayesian model averaged" strategy exceeds that under the strategy that conditions on G, apparently naturally induced by diversity in some aspects of the underlying graphical model structure that induces more variation in portfolio weights. As with the modal graph G, the model averaged graph beats the full graph in the sense of having smaller risk for a fixed target return, as well as representing inferences based on graphs that very substantially fit the data better than the full graph. Hence, whether based on posterior modes over graphs or model averaging over graph structure, the utility of posterior simulation in the structured HIW framework is evident.

4.2 Discussion

This chapter has presented and illustrated direct simulation methods for hyper-inverse Wishart distributions on Gaussian graphical models. The methods apply to both decomposable and non-decomposable models, and takes advantage

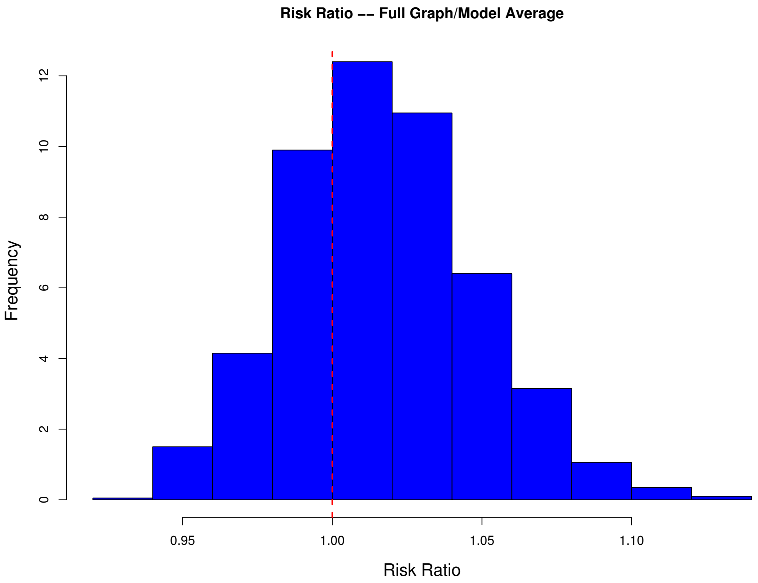


Figure 4.6: Posterior distribution of the ratio of standard deviations of the optimal portfolios under the full graph relative to that under model average.

of the junction tree representation to efficiently sample a complete but possibly very highly structured variance matrix via composition. One attractive feature of the methods is the fact that inversions and decompositions of large matrices are avoided. The methods require matrix manipulations and Cholesky decompositions for simulation that involve matrices of no higher dimension than the size of the largest prime component of the graph. This is a critical advantage and a key for applications in higher dimensional problems. The examples presented were of moderate size however the efficiency of the method is maintained in larger problems. I have experimented with the approach in problems up to 200 nodes and computations remain very efficient. To illustrate how computations scale up a simulation study was performed. For different dimensions (10, 30, 50, 100 and 150) two sets of 10 decomposable graphs were simulated from the MH algorithm

Upper Line					
Nodes	10	30	50	100	150
Number of Cliques	7	27	47	89	130
Number of Edges	11	31	51	118	204

Lower Line					
Nodes	10	30	50	100	150
Number of Cliques	2	18	30	70	99
Number of Edges	25	89	160	276	572

Table 4.2: Structure of simulated graphs for cpu benchmark studies. The table gives the median number of cliques and edges in the 100 generated graphs for each case (number of nodes) under the two different sparsity priors - the upper (red) and lower (blue) lines examples in Figure 4.7.

presented in Chapter 3. Evidently, computation time increases with the complexity of the graph, in terms of the size of larger cliques in decomposable graphs and the nature and dimension of larger prime components in non-decomposable cases. The two sets differ in the sparsity prior used; For the first set $\beta = 2/(149)$ whereas in the second $\beta = 4/(149)$. The idea here is to compare the efficiency of the methods in graphs the not only differ by size but also by complexity. Table 4.2 displays some of the median characteristics of the graphs in each set. The data used was that of the 150 node gene expression example of section 3.6. Figure 4.7 shows the results of the simulation and indicates that computations increase roughly linearly with dimension. This is understandable as the number of cliques in large decomposable graphs with similar degrees of sparsity will increase linearly and so does the computational burden. The situation is similar with non-decomposable graphs, though understanding scalability there requires further study, since the complexity of non-decomposable graphs is impacted by the complexity of the structure of its prime components as well as just the distribution of component size.

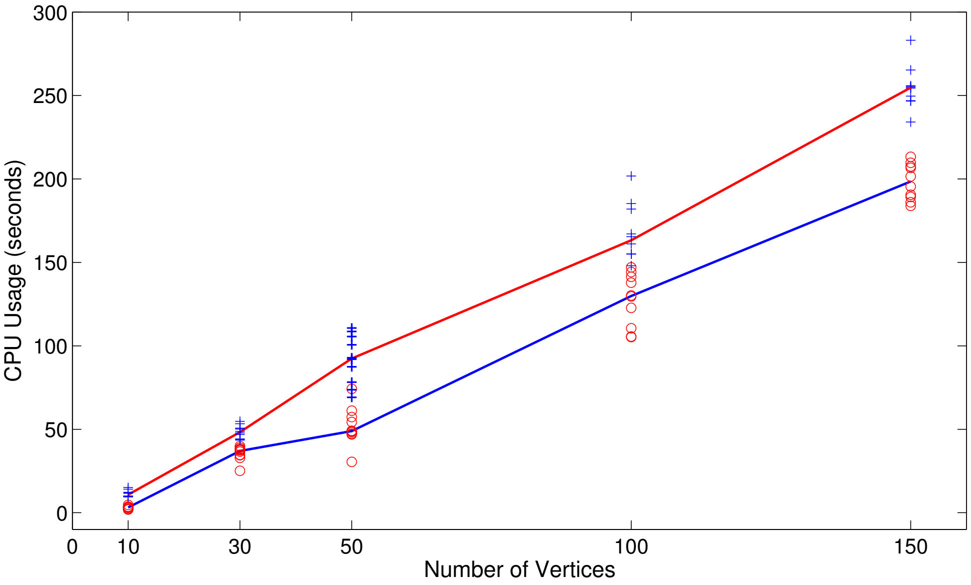


Figure 4.7: Compute time as a function of size of graph. The graph shows the increase in cpu time to simulate the HIW distribution 100 times on a decomposable graph, and how the time changes as a function of the dimension (number of vertices). Graphs were generated randomly and the upper (red) and lower (blue) lines represent differing degrees of sparsity: the upper cases correspond to graphs in which edges occur with probability $2/p$, and the lower those with probability $4/p$, where p is the number of vertices. The crosses (relate to the red line) and circles (relate to the blue line) represent cpu times for specific simulated graphs.

Chapter 5

Matrix DLMS with Graphical Model Structuring

In this chapter I introduce a new class of models for multivariate time series analysis based on the general idea of sparsity in modeling covariance structures. The models establish a connection between multivariate dynamic linear models (DLM) and graphical models through the use of the hyper-inverse Wishart (HIW) distribution as a prior for the cross sectional covariance structure of p time series.

The chapter starts with some background on multivariate DLMs, followed by the generalization to HIW case. Time-varying covariances and model uncertainty are then discussed. A comprehensive illustration of the methodology is presented *via* two financial time series portfolio selection examples.

5.1 The Multivariate Normal DLM

Bayesian dynamic linear models (DLM) represent a broad class of structural forecasting models that have been extensively used in many application areas such as finance, engineering, ecology and medicine. In general, DLMs are dynamic linear regression models (or state-space models) with Markovian evolution structures. The sequential model specification and flexibility allows for the creation of complex forecasting models where expert information and systematic interventions can easily be incorporated. A thorough discussion of DLMs can be found in West and Harrison (1997).

The focus of this chapter is on a specific (and yet general) subclass of DLMs, the Matrix Normal DLMs as presented in Quintana (1987) and applied in Quintana and West (1987). These models create a general, fully-conjugate, framework for multivariate time series analysis when the cross sectional covariance matrix is not known. In order to provide close-form analytical solutions, the model requires common components defining each individual DLM, thus making these models suited for the analysis of similar time series such as stock prices, bond prices, temporal gene expression data, etc.

Specify the model via individual univariate components. Let p univariate time series Y_{ti} following individual DLMs defined by

$$\left\{\mathbf{F}_t, \mathbf{G}_t, V_t\sigma_i^2, \mathbf{W}_t\sigma_i^2\right\}.$$

Here, t is the time indicator whereas i indexes each time series ($i = 1, \ldots, p$). All quantities are assumed known aside from σ_i^2, $\forall i = 1, \ldots, p$. Each univariate model can be written as:

$$\begin{aligned} \text{Observation:} \quad Y_{ti} &= \mathbf{F}_t\boldsymbol{\theta}_{ti} + \nu_{ti}, \qquad && \nu_{ti} \sim N(0, V_t\sigma_i^2) && (5.1) \\ \text{Evolution:} \quad \boldsymbol{\theta}_{ti} &= \mathbf{G}_t\boldsymbol{\theta}_{t-1,i} + \boldsymbol{\omega}_{ti}, \qquad && \boldsymbol{\omega}_{ti} \sim N(0, \mathbf{W}_t\sigma_i^2) && (5.2) \end{aligned}$$

Some standard conditional independence assumptions are necessary: given all parameters the random innovations ν_{ti} and $\boldsymbol{\omega}_{ti}$ are independent across time and mutually independent.

To complete the model, a cross sectional covariance structure ($\mathbf{\Sigma}$) is defined through covariances between the observational and evolution errors.

Stack the individual DLMs and establish the following notation (at time t):

- $\mathbf{Y}_t = (Y_{t1}, \ldots, Y_{tp})'$, a $p \times 1$ vector;
- $\mathbf{\Theta}_t = (\boldsymbol{\theta}_{t1}, \ldots, \boldsymbol{\theta}_{tp})$, a $n \times p$ matrix of states;
- $\mathbf{\Omega}_t = (\boldsymbol{\omega}_{t1}, \ldots, \boldsymbol{\omega}_{tp})$, a $n \times p$ matrix of evolution innovations;
- $\boldsymbol{\nu}_t = (\nu_{t1}, \ldots, \nu_{tp})'$, a $p \times 1$ vector of observational innovations.

The full model can then be stated as

$$\mathbf{Y}'_t = \mathbf{F}'_t \mathbf{\Theta}_t + \boldsymbol{\nu}'_t, \qquad \boldsymbol{\nu}_t \sim N(\mathbf{0}, V_t \mathbf{\Sigma}) \tag{5.3}$$

$$\mathbf{\Theta}_t = \mathbf{G}_t \mathbf{\Theta}_{t-1} + \mathbf{\Omega}_t \qquad \mathbf{\Omega}_t \sim N(\mathbf{0}, \mathbf{W}_t, \mathbf{\Sigma}) \tag{5.4}$$

where the evolution innovation matrix $\mathbf{\Omega}_t$ follows a *matrix-variate normal* distribution as defined by Dawid (1981) (details in Appendix) with mean $\mathbf{0}$ (a $n \times p$ matrix), left covariance matrix $\mathbf{W}_t$ (rows) and right covariance matrix $\mathbf{\Sigma}$ (columns).

The cross-sectional structure across series comes in via the elements σ_{ij} ($i, j = 1, \ldots, p$) of the ($p \times p$) covariance matrix $\mathbf{\Sigma}$. The model, as described in (5.3) and (5.4), implies that (for all $i, j = 1, \ldots, p$)

$$Cov(\nu_{ti}, \nu_{tj}) = V_t \sigma_{ij}$$

$$Cov(\boldsymbol{\omega}_{ti}, \boldsymbol{\omega}_{tj}) = \mathbf{W}_t \sigma_{ij}.$$

The correlation structure induced by $\mathbf{\Sigma}$ affects both the observational and evolution errors thus if σ_{ij} is large and positive, series i and j will show a similar behavior in both their underlying state evolution and in the observational variation about their level.

5.1.1 Updating Recurrences

Based on a normal/inverse-Wishart prior for the state matrix $\mathbf{\Theta}$ and the covariance matrix $\mathbf{\Sigma}$, a conjugate, sequential updating estimation procedure is available. Suppose that the initial prior for $\mathbf{\Theta}_0$ and $\mathbf{\Sigma}$ is a normal/inverse Wishart (NIW hereafter) denoted by

$$(\mathbf{\Theta}_0, \mathbf{\Sigma}|D_0) \sim NIW(\mathbf{m}_0, \mathbf{C}_0, b_0, \mathbf{S}_0) \tag{5.5}$$

where $\mathbf{m}_0$,$\mathbf{C}_0$,$\mathbf{S}_0$ and b_0 are pre-defined quantities and D_0 is the information set at time 0. At each step, the information set is updated so that $D_t = \{Y_t, D_{t-1}\}$. The notation in Equation (5.5) implies that

$$(\mathbf{\Theta}_0|\mathbf{\Sigma}, D_0) \sim N(\mathbf{m}_0, \mathbf{C}_0, \mathbf{\Sigma}) \quad \text{and} \quad (\mathbf{\Sigma}|D_0) \sim IW(b_0, \mathbf{S}_0).$$

For all $t > 1$ the following theorem apply (as described in West and Harrison (1997) and proved in detail by Quintana (1987)).

Theorem 5.1. *The sequential updating for the matrix normal DLM is given as follows:*

(i) *Posterior at $t-1$:*

$$(\mathbf{\Theta}_{t-1}, \mathbf{\Sigma}|D_{t-1}) \sim NIW(\mathbf{m}_{t-1}, \mathbf{C}_{t-1}, b_{t-1}, \mathbf{S}_{t-1})$$

(ii) *Prior at t:*

$$(\mathbf{\Theta}_t, \mathbf{\Sigma}|D_{t-1}) \sim NIW(\mathbf{a}_t, \mathbf{R}_t, b_{t-1}, \mathbf{S}_{t-1})$$

where

$$\mathbf{a}_t = \mathbf{G}_t\mathbf{m}_{t-1} \quad \textit{and} \quad \mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t' + \mathbf{W}_t$$

(iii) *One-step forecast:*

$$(\mathbf{Y}_t|D_{t-1}) \sim T(\mathbf{f}_t, Q_t\mathbf{S_{t-1}}, b_{t-1})$$

where

$$\mathbf{f}'_t = \mathbf{F}'_t\mathbf{a}_t \quad \textit{and} \quad Q_t = \mathbf{F}'_t\mathbf{R}_t\mathbf{F}_t + V_t$$

(iv) *Posterior at t:*

$$(\mathbf{\Theta}_t, \mathbf{\Sigma}|D_t) \sim NIW(\mathbf{m}_t, \mathbf{C}_t, b_t, \mathbf{S}_t)$$

with

$$\begin{aligned}
\mathbf{m}_t &= \mathbf{a}_t + \mathbf{A}_t\mathbf{e}'_t \\
\mathbf{C}_t &= \mathbf{R}_t - \mathbf{A}_t\mathbf{A}'Q_t \\
b_t &= b_{t-1} + 1 \\
\mathbf{S}_t &= \mathbf{S}_{t-1} + \mathbf{e}_t\mathbf{e}'_t/Q_t
\end{aligned}$$

where

$$\mathbf{A}_t = \mathbf{R}_t\mathbf{F}_t/Q_t \quad \textit{and} \quad \mathbf{e}_t = \mathbf{Y}_t - \mathbf{f}_t$$

Proof. See Quintana (1987) □

It is important to emphasize that due to the common components $\mathbf{F}_t$,$\mathbf{G}_t$ and $\mathbf{W}_t$, fitting a matrix normal DLM is almost equivalent to fitting p individual DLMs to each of the series. The difference lies on the ability to estimate the cross sectional covariance structure $\mathbf{\Sigma}$, where its marginal posterior distribution is given by

$$(\mathbf{\Sigma}|D_t) \sim IW(b_t, \mathbf{S}_t).$$

This class of models has proved very useful in dynamic portfolio problems (Quintana and West, 1987; Quintana, 1992; Quintana *et al.*, 2003) where sequential investment decisions are made based on estimates of returns, volatility and co-variation between assets. As p increases the estimation of $\mathbf{\Sigma}$ becomes more difficult with increasing instability of optimal portfolio weights (Ledoit and Wolf, 2004). Parsimonious methods that allow for a reduction in the parameter space are of interest and graphical model structure arises as a natural way of address this problem.

5.2 Sparsity in DLMs: Generalization to HIW

As discussed in Chapters 2 and 3, Gaussian graphical models are a representation of conditional independence structure in multivariate distributions where decompositions of the joint distribution provide computational efficiencies and a reduction in the space of parameters. Taking advantage of the latter, this section shows how graphical structuring can be incorporated in matrix normal DLMs providing a parsimonious model for $\mathbf{\Sigma}$. For a given decomposable graph, the hyper-inverse Wishart is used as a conjugate prior for $\mathbf{\Sigma}$ and Theorem 5.1 is generalized, with the analytical, closed-form, sequential updating procedure being preserved.

Consider the matrix normal DLM as described in Equations (5.3) and (5.4). Now, the conjugate prior for $\mathbf{\Theta}$ and $\mathbf{\Sigma}$ is a normal/hyper-inverse Wishart (as defined in Chapter 2) denoted by

$$(\mathbf{\Theta}_0, \mathbf{\Sigma} | D_0) \sim NHIW_G(\mathbf{m}_0, \mathbf{C}_0, b_0, \mathbf{S}_0),$$

meaning that

$$(\mathbf{\Theta}_0 | \mathbf{\Sigma}, D_0) \sim N(\mathbf{m}_0, \mathbf{C}_0, \mathbf{\Sigma}) \qquad \text{and} \qquad (\mathbf{\Sigma} | D_0) \sim HIW_G(b_0, \mathbf{S}_0). \tag{5.6}$$

Updating recurrences are presented in the following theorem.

Theorem 5.2. *The sequential updating for the matrix normal DLM with graphical structure is given as follows:*

(i) *Posterior at $t-1$:*

$$(\mathbf{\Theta}_{t-1}, \mathbf{\Sigma}|D_{t-1}) \sim NHIW_G(\mathbf{m}_{t-1}, \mathbf{C}_{t-1}, b_{t-1}, \mathbf{S}_{t-1})$$

(ii) *Prior at t:*

$$(\mathbf{\Theta}_t, \mathbf{\Sigma}|D_{t-1}) \sim NHIW_G(\mathbf{a}_t, \mathbf{R}_t, b_{t-1}, \mathbf{S}_{t-1})$$

where

$$\mathbf{a}_t = \mathbf{G}_t\mathbf{m}_{t-1} \quad \textit{and} \quad \mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}'_t + \mathbf{W}_t$$

(iii) *One-step forecast:*

$$(\mathbf{Y}_t|D_{t-1}) \sim HT_G(\mathbf{f}_t, Q_t\mathbf{S_{t-1}}, b_{t-1})$$

where

$$\mathbf{f}'_t = \mathbf{F}'_t\mathbf{a}_t \quad \textit{and} \quad Q_t = \mathbf{F}'_t\mathbf{R}_t\mathbf{F}_t + V_t$$

(iv) *Posterior at t:*

$$(\mathbf{\Theta}_t, \mathbf{\Sigma}|D_t) \sim NHIW_G(\mathbf{m}_t, \mathbf{C}_t, b_t, \mathbf{S}_t)$$

with

$$\begin{aligned} \mathbf{m}_t &= \mathbf{a}_t + \mathbf{A}_t\mathbf{e}'_t \\ \mathbf{C}_t &= \mathbf{R}_t - \mathbf{A}_t\mathbf{A}'Q_t \\ b_t &= b_{t-1} + 1 \\ \mathbf{S}_t &= \mathbf{S}_{t-1} + \mathbf{e}_t\mathbf{e}'_t/Q_t \end{aligned}$$

where

$$\mathbf{A}_t = \mathbf{R}_t\mathbf{F}_t/Q_t \quad \textit{and} \quad \mathbf{e}_t = \mathbf{Y}_t - \mathbf{f}_t$$

Proof. This theorem is a direct extension of Theorem 5.1. The results that require a different proof are *(iii)* and the updating related to $\mathbf{\Sigma}$ in *(iv)*.

- *Proof of (iii)*: It is clear that

  $$(\mathbf{Y}_t|\mathbf{\Sigma}, D_{t-1}) \sim N(\mathbf{f}_t, Q_t\mathbf{\Sigma}),$$

  with $(\mathbf{\Sigma}|D_{t-1}) \sim HIW_G(b_{t-1}, S_{t-1})$ so, for each clique C, the marginal distribution of $\mathbf{Y}_t^C$ is simply a $T(\mathbf{f}_t, Q_t\mathbf{S}_{t-1}^C, b_{t-1})$. The overall marginal distribution of $\mathbf{Y}_t$ is then a hyper-T distribution given by the Markov combination (consistent with G) of T-distributions over cliques and separators, as defined in Dawid and Lauritzen (1993), and denoted here by HT_G (see Chapter 2).

- *Proof of (iv)*: The updating for $\mathbf{\Sigma}$ follows directly the conjugacy results for the HIW described in Chapter 2 (Equations 2.12 and 2.13).

□

5.3 Retrospective Recurrences

After observing a fixed set $\mathbf{Y}_{1:T} = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_T\}$ one might be interested in looking back in time to more clearly understand what happened. The uncertainty about the state matrix at time t is now updated in light of all observations up to time T. Now the interest lies on $p(\mathbf{\Theta}_{T-k}|D_T)$ for all $k \geq 1$, the so-called k-step filtered distribution for the state matrix.

Using recent data to revise inferences about previous values of the state vector is simply implemented as a direct byproduct of conditional independences of DLMs and, given that $\mathbf{\Sigma}$ is a fixed parameter (not a state), the results developed in West and Harrison (1997) extend to the matrix DLMs with graphical structure.

In summary, the filtered distribution of the state matrix $\mathbf{\Theta}_{t-k}$ and $\mathbf{\Sigma}$ is defined as (details in West and Harrison, 1997):

$$(\mathbf{\Theta}_{t-k}, \mathbf{\Sigma}|D_t) \sim NHIW_G(\mathbf{a}(-k)_t, \mathbf{R}(-k)_t, b_t, \mathbf{S}_t) \tag{5.7}$$

where the parameters are calculated through the following recurrences:

$$\begin{aligned}
\mathbf{B}_{t-k} &= \mathbf{C}_{t-k}\mathbf{G}'_{t-k+1}\mathbf{R}^{-1}_{t-k+1} \\
\mathbf{a}_t(-k) &= \mathbf{m}_{t-k} + \mathbf{B}_{t-k}[\mathbf{a}_t(-k+1) - \mathbf{a}_{t-k+1}] \\
\mathbf{R}_t(-k) &= \mathbf{C}_{t-k} + \mathbf{B}_{t-k}[\mathbf{R}_t(-k+1) - \mathbf{R}_{t-k+1}]\mathbf{B}'_{t-k},
\end{aligned}$$

with starting values

$$\mathbf{a}_t(0) = \mathbf{m}_t \qquad \text{and} \qquad \mathbf{R}_t(0) = \mathbf{C}_t.$$

5.4 Time-Varying $\mathbf{\Sigma}$

So far, $\mathbf{\Sigma}$ has been held fixed, not varying through time. The possibility of $\mathbf{\Sigma}$ varying stochastically over time is very attractive, especially in financial applications where ARCH type models (Bollerslev *et al.*, 1992) of conditional variance and stochastic volatility models (Jacquier *et al.*, 1994; Kim *et al.*, 1998) are very popular. Variance discounting ideas (Ameen and Harrison, 1985; West and Harrison, 1997; Quintana *et al.*, 2003) can be adapted to matrix normal DLMs creating an evolution process that decay the information about $\mathbf{\Sigma}$ between time points while maintaining the nice conjugate results of Theorem 5.2. The degree of information loss is determined by the variance discount factor δ with $0 < \delta \leq 1$. This parameter controls how adaptive to new information the model will be. Values of δ close to 1 imply a model that preserves a lot of information through time and current estimates of $\mathbf{\Sigma}$ are heavily dependent on the far past. In the limit, with $\delta = 1$, the model is static and is back to a fixed covariance structure. The opposite occurs for small values of δ in which case the model adapts to new information very fast and estimates will be heavily concentrated on recent information.

First, suppose that G is the full graph and the posterior for $\mathbf{\Sigma}_{t-1}$ at time $t-1$ is given by

$$(\mathbf{\Sigma}_{t-1}|D_{t-1}) \sim IW(b_{t-1}, \mathbf{S}_{t-1}).$$

Note that the time dependency is now explicit so that the covariance structure at time $t-1$ is $\mathbf{\Sigma}_{t-1}$. Evolving to time t we want to maintain the inverse-Wishart form for the prior of $\mathbf{\Sigma}_t$ with, possibly the same location, but with an increased dispersion due to the loss of information from $t-1$ to t. This is accomplished by

$$(\mathbf{\Sigma}_t|D_{t-1}) \sim IW(\delta b_{t-1}, \delta\mathbf{S}_{t-1}) \tag{5.8}$$

where the increment in dispersion derives from the fact that $\delta b_{t-1} < b_{t-1}$. This evolution is a direct analogue to discount ideas in DLMs (West and Harrison, 1997) and formal models underlying this process are described in Uhlig (1994) and Quintana *et al.* (1995). The general concept is based on the combination of Wisharts and a matrix-Beta distribution in a multiplicative model that creates a "random walk" model for $\mathbf{\Sigma}_t^{-1}$. In short, write $\mathbf{\Sigma}_t^{-1} = \mathbf{U}'_{t-1}\mathbf{B}_t\mathbf{U}_{t-1}$ where $\mathbf{\Sigma}_{t-1}^{-1} = \mathbf{U}'_{t-1}\mathbf{U}_{t-1}$. Given that $(\mathbf{\Sigma}_{t-1}^{-1}|D_{t-1})$ is a Wishart random variable, choosing the appropriate form of matrix-Beta for $\mathbf{B}_t$ will deliver the desired prior for $(\mathbf{\Sigma}_t^{-1}|D_{t-1})$. The drawback of the idea above is the lack of flexibility in discounting. The appropriate choice of hyper-parameters for the distribution of $\mathbf{B}_t$ restricts the discount factor to unusual values, outside of the standard $0.75 \leq \delta < 1$ interval. An alternative, more flexible proposal appears in Liu (2000) and Quintana *et al.*

(2003) where the map from $\mathbf{\Sigma}_{t-1}$ to $\mathbf{\Sigma}_t$ is establish through multiplicative beta shocks applied to the diagonal of the Bartlett decomposition of $\mathbf{\Sigma}_{t-1}^{-1}$. Adapting the latter evolution to a situation where G is not full, consider, without loss of generality, the following DLM:

$$\mathbf{Y}_t \sim N(\mathbf{0}, \mathbf{\Sigma}_t) \tag{5.9}$$

with posterior at time $t-1$ given by

$$(\mathbf{\Sigma}_{t-1}|D_{t-1}) \sim HIW_G(b_{t-1}, \mathbf{S}_{t-1}).$$

Following the hyper-inverse Wishart extension of the Bartlett decomposition (Roverato, 2000) set $\mathbf{\Sigma}_{t-1} = \mathbf{\Phi}^{-1'}\mathbf{\Phi}^{-1}$ where:

$$\mathbf{S}_{t-1}^{-1} = \mathbf{T}'\mathbf{T} \qquad \text{and} \qquad \mathbf{\Phi} = \mathbf{\Psi}\mathbf{T}$$

with $\mathbf{\Psi} \in M^t(G)$, and

$$(\psi_{ii})^2 \sim \chi^2_{b_{t-1}+\nu_i} \qquad \text{and} \qquad \psi_{ij} \sim N(0,1)$$

for $(i,j) \in V$.

Here, $M^t(G)$ is the set of all upper triangular matrices with positive diagonal elements such that the entries $(i,j) \notin E$ are zero, and ν_i is the number of non-zero elements in row i of $\mathbf{\Psi}$. Evolving the posterior at time $t-1$ to the prior at time t follows a transformation of the diagonal elements of $\mathbf{\Psi}$ where for $\mathbf{\Psi}^* \in M^t(G)$:

$$\psi^*_{ii} = \psi_{ii}\sqrt{r_i}$$

and

$$\psi^*_{ij} = \psi_{ij} \quad \text{for} \quad (i,j) \in V,$$

with

$$r_i \sim Beta\left[\frac{\delta_i}{2}(b_{t-1}+\nu_i), \frac{(1-\delta_i)}{2}(b_{t-1}+\nu_i)\right]$$

and

$$\delta_i = \frac{\delta b_{t-1}+\nu_i}{b_{t-1}+\nu_i}.$$

Now,

$$(\psi^*_{ii})^2 \sim \chi^2_{\delta b_{t-1}+\nu_i}$$

and

$$\psi^*_{ij} \sim N(0,1) \quad \text{for} \quad (i,j) \in V,$$

so that by setting

$$\begin{aligned} \mathbf{T}^* &= \sqrt{\delta}\mathbf{T}, \\ \mathbf{\Phi}^* &= \mathbf{\Psi}^*\mathbf{T}^* \end{aligned}$$

and

$$\mathbf{\Sigma}_t = \mathbf{\Phi}^{*-1\prime}\mathbf{\Phi}^{*-1},$$

the prior for $\mathbf{\Sigma}_t$ takes the form

$$(\mathbf{\Sigma}_t|D_{t-1}) \sim HIW_G(\delta b_{t-1}, \delta\mathbf{S}_{t-1}). \tag{5.10}$$

By using the above prior, updating is straightforward with the posterior at time t given by

$$(\mathbf{\Sigma}_t|D_t) \sim HIW_G(b_t, \mathbf{S}_t)$$

with

$$b_t = \delta b_{t-1} + 1 \quad \text{and} \quad \mathbf{S}_t = \delta \mathbf{S}_{t-1} + \mathbf{Y}_t \mathbf{Y}_t'. \tag{5.11}$$

In this model, as $t \longrightarrow \infty$ the posterior mean of $\mathbf{\Sigma}_t$ becomes an exponentially weighted moving average estimate of the covariance structure. This follows since $b_t \longrightarrow (1-\delta)^{-1}$ and $E(\mathbf{\Sigma}_t^C | D_t) = \mathbf{S}_t^C/(b_t - 2)$ so that

$$E(\mathbf{\Sigma}_t^C | D_t) \approx (1-\delta) \sum_{l=0}^{t-1} \delta^l \mathbf{Y}_{t-l}^C \mathbf{Y}_{t-l}^{C\prime}.$$

This provides a framework where forward estimates of $\mathbf{\Sigma}_t$ keep adapting to new data while further discounting past observations. Modifications in the discount factor also allow for abrupt changes in volatility to be incorporated into the model giving the forecaster the ability to determine how fast the model should react to external information.

As discussed in West and Harrison (1997) no complete closed form retrospective update is available for $\mathbf{\Sigma}_t$. However the mean of the filtered distribution of $\mathbf{\Sigma}_t^{-1}$ (under the full graph) can be recursively calculated by

$$E(\mathbf{\Sigma}_{t-k}^{-1} | D_t) = \mathbf{S}_t^{-1}(-k)[b_t(-k) + p - 1] \tag{5.12}$$

where,

$$\begin{aligned} \mathbf{S}_t^{-1}(-k) &= (1-\delta)\mathbf{S}_{t-k}^{-1} + \delta \mathbf{S}_t(-k+1)^{-1} && (5.13) \\ b_t(-k) &= (1-\delta) b_{t-k} + \delta b_t(-k+1) && (5.14) \end{aligned}$$

with starting values

$$\mathbf{S}_t(0) = \mathbf{S}_t \qquad \text{and} \qquad b_t(0) = b_t.$$

So, for every clique $C \in \mathcal{C}$ (and separator $S \in \mathcal{S}$), retrospective estimates of $\mathbf{\Omega}_t^C$ can be computed by (5.12) and if combined with (2.9) provide retrospective estimates for $\mathbf{\Omega}_t$.

5.5 Large-Scale Dynamic Portfolio Allocation

The use of Bayesian forecasting through DLMs in asset allocation problems has been routine for a number of years. The one-step ahead forecast distribution of future returns is the key component for mean-variance portfolio optimization that allow for parameter uncertainty to be taken into account in sequential investment decisions. Aspects of Bayesian portfolio selection are discussed in detail in Polson and Tew (2000); Quintana (1992), Putnam and Quintana (1994), Quintana and Putnam (1996) and Quintana *et al.* (2003) are examples of carefully developed DLMs that implement portfolio rules in fixed income and currency markets.

The static portfolio example of Chapter 4 is a first illustration of how sparsity in modeling the covariance of assets can reduce the uncertainty about optimal portfolio weights and so induce less volatile investment opportunities. DLMs with sparse covariance matrices, as developed in Theorem 5.2, are a way to revisit and explore that simple example in a more realistic fashion, i.e. in a dynamic allocation process. Throughout the applications presented in this section, optimal portfolios weights ($\mathbf{w}_t$) are determined based on the quadratic programming procedures developed by Markowitz (1959). In this set-up, given the first two moments of the predictive distribution of returns, say $\mathbf{f}_t$ and $\mathbf{Q}_t$, and a fixed return target m, the investor decision problem reduces to minimizing the one-step ahead portfolio variance $\mathbf{w}_t'\mathbf{Q}_t^{-1}\mathbf{w}_t$ subject to constraints $\mathbf{w}_t'\mathbf{f}_t = m$ and $\mathbf{w}_t'\mathbf{1} = 1$. The general solution for the above optimization through Lagrange multipliers creates the so called *efficient frontier* where the mean-variance efficient portfolio is given by

$$\mathbf{w}_t^{(m)} = \mathbf{Q}_t^{-1}(a\mathbf{f}_t + b\mathbf{1}) \tag{5.15}$$

where

$$a = \mathbf{1}'\mathbf{Q}_t^{-1}\mathbf{e} \quad \text{and} \quad b = -\mathbf{f}_t'\mathbf{Q}_t^{-1}\mathbf{e}$$

with

$$\mathbf{e} = \frac{(\mathbf{1}m - \mathbf{f}_t)}{d} \quad \text{and} \quad d = (\mathbf{1}'\mathbf{Q}_t^{-1}\mathbf{1})(\mathbf{f}_t'\mathbf{Q}_t^{-1}\mathbf{f}_t) - (\mathbf{1}'\mathbf{Q}_t^{-1}\mathbf{f}_t)^2.$$

An interesting alternative portfolio that involves only estimates of $\mathbf{Q}_t$ is the minimum-variance portfolio where,

$$\mathbf{w}_t = \frac{\mathbf{Q}_t^{-1}\mathbf{1}}{\mathbf{1}'\mathbf{Q}_t^{-1}\mathbf{1}}. \tag{5.16}$$

This strategy isolates the effects of $\mathbf{Q}_t$ on investment decisions and is of interest when competing models for covariance estimation are considered.

Perold (1988), Polson and Tew (2000) and Ledoit and Wolf (2004) point out that building high-dimensional portfolios tend to result in extreme and very unstable weights assigned to each asset. This is due to the large amount of uncertainty in the estimation of covariance matrices, especially when the the number of historical observations is relatively small if compared to the number of assets considered. From (5.15) and (5.16) it is clear that the solution for optimal portfolios is a direct function of the precision matrix $\mathbf{K}_t = \mathbf{Q}_t^{-1}$. A nice representation appears in Stevens (1998), namely

$$\mathbf{w}_{it}^{(m)} = \lambda \frac{f_{it} - \sum_{j \neq i} (k_{ij}/k_{ii}) f_{jt}}{k_{ii}^{-1}} \tag{5.17}$$

with λ being the Lagrange multiplier. If it is assumed that the returns are normally distributed, expression (5.17) shows that the weight assigned to asset i depends on the ratio of the intercept of its regression on all other assets relative to the conditional variance of the regression. In other words, the amount of money invested in asset i depends on the ratio of the expected return that cannot be explained by the linear combination of assets over the *unhedgeable* (or *nondiversifiable*) risk.

Note that the numerator is a function of the off-diagonal elements of $\mathbf{K}_t$ hence it is not surprising that conditional independence assumptions have a direct influence over the uncertainty about $\mathbf{w}_t$. If, in fact, the unhedgeable risk can be obtained by a regression involving a smaller number of regressors (i.e. having some of the k_{ij}'s equal to zero) this has to be taken into account; failing to do so implies that unnecessary parameters are being estimated and nothing but uncertainty is added to the problem. In the following two applications, I show how imposing conditional independence constraints help create portfolios that not only are less risky but also turn out to be more profitable.

First the exchange rate data of Chapter 4 is revisited. This is followed by an example involving a large set of securities in the S&P 500 stock index. The goal is to simply compare the performance of dynamic portfolios built from both a "full" (unconstrained) and a "sparse" (with graphical constraints) DLM. In all examples a simple DLM with time-varying covariance structure (as in Section 5.4) is used, namely

$$\mathbf{Y}_t = \boldsymbol{\theta}_t + \boldsymbol{\nu}_t \qquad \boldsymbol{\nu}_t \sim N(0, \boldsymbol{\Sigma}_t) \tag{5.18}$$

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t \quad \boldsymbol{\omega}_t \sim N(0, W_t\boldsymbol{\Sigma}_t). \tag{5.19}$$

This is special case of the general DLMs presented in previous sections and updating recurrences are available from Theorem 5.1 and Theorem 5.2 along with Equations (5.11).

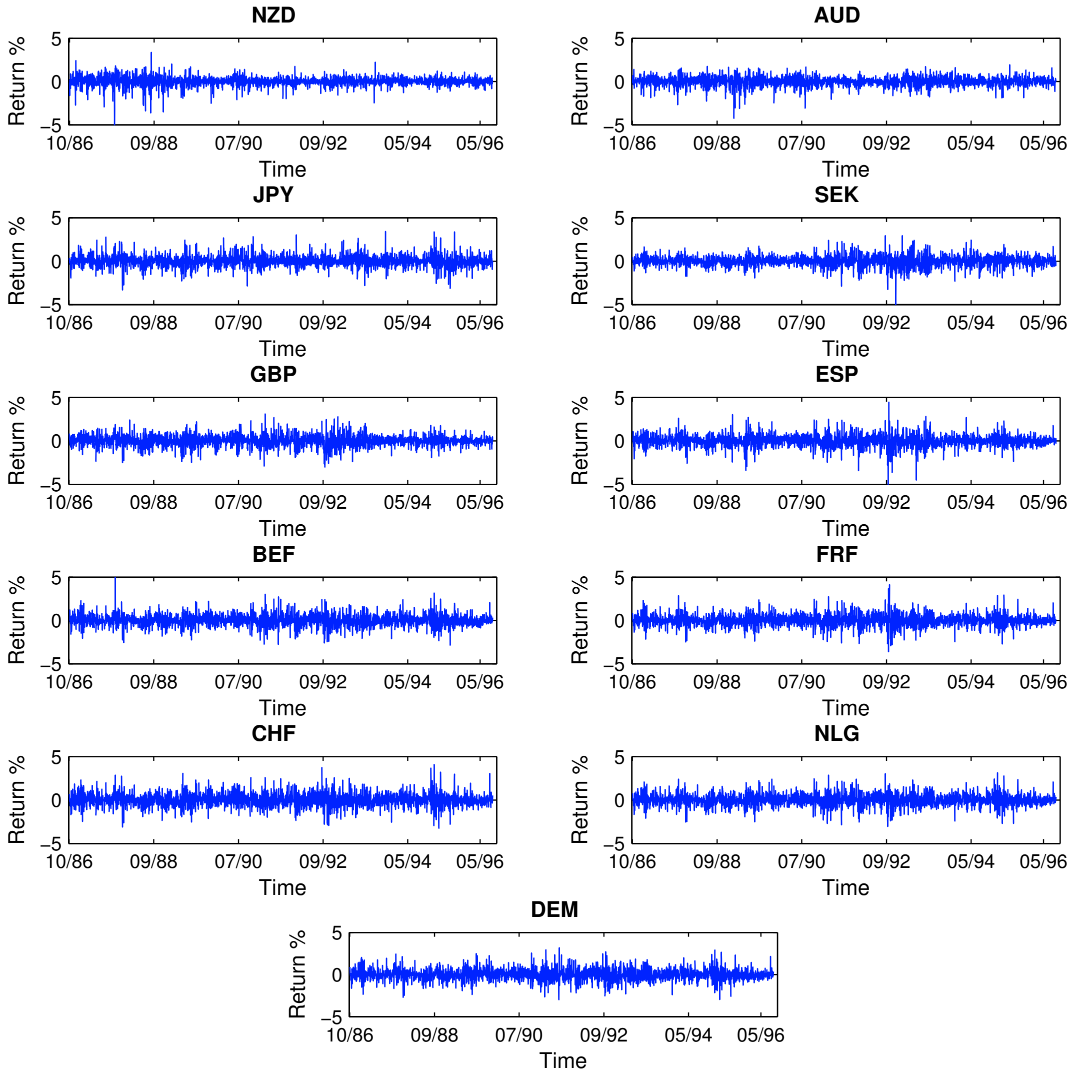
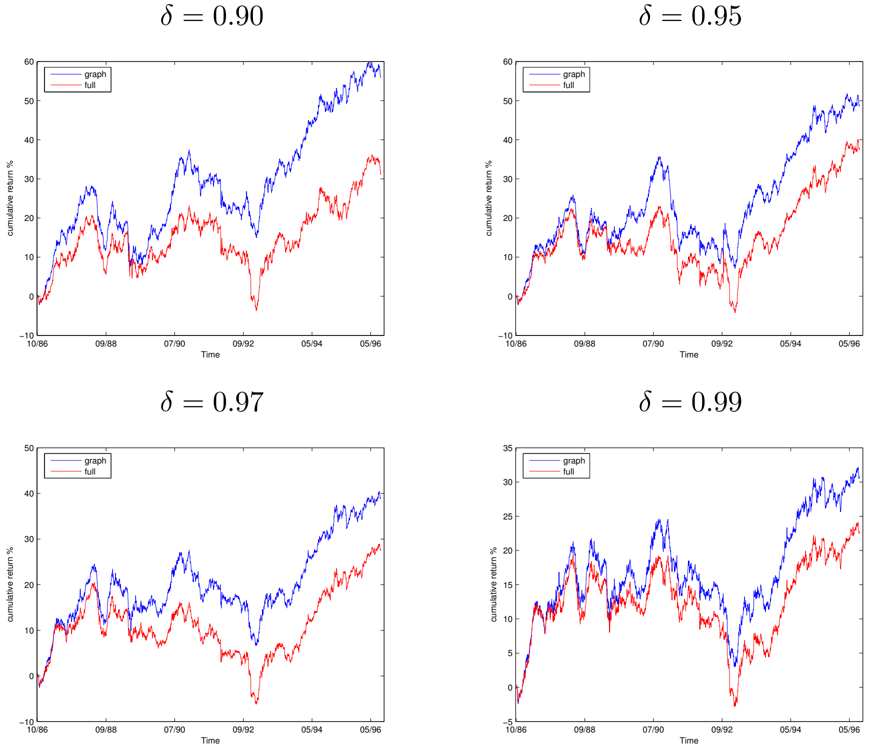


Figure 5.1: Daily exchange rate returns. The dataset consists of 2566 daily returns for each of these 11 currencies, over the period of about 10 years – 10/9/86 to 8/9/96



$\delta = 0.90$

$\delta = 0.95$

$\delta = 0.97$

$\delta = 0.99$

Figure 5.2: Cumulated returns for mean-target portfolios in the unconstrained ("full") and sparse ("graph") model. Portfolio weights were computed using the $E(\mathbf{Y}_t|D_{t-1})$ and $Var(\mathbf{Y}_t|D_{t-1})$.

5.5.1 Dynamic Portfolios for International Exchange Rates

Consider the exchange rate data used in Example 4.1.3 and displayed in Figure 5.1. The graph for the constrained model is also the same as in Chapter 4 (Figure 4.1). In all models, non-informative priors (very small b_0) with scale matrix equal to the identity are used and four separate analysis are presented with different discount factor δ taking the values 0.90, 0.95, 0.97 and 0.99. Most of the discussion is focused on the model with $\delta = 0.97$. For the purpose of this example δ is not important and different values are considered just to show that imposing graphical structures helps in all cases.

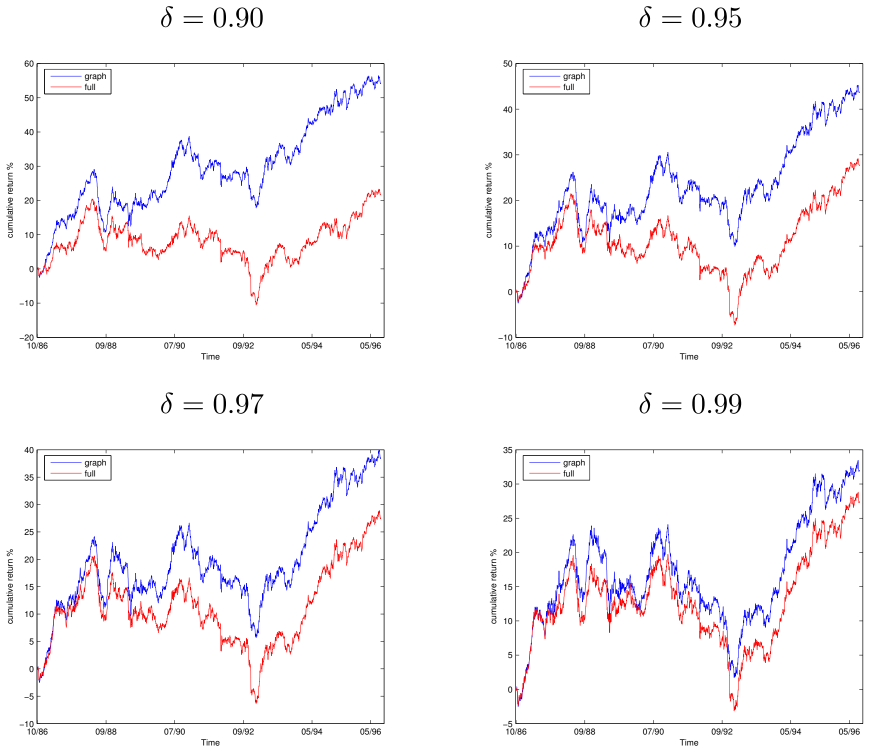


Figure 5.3: Cumulated returns for minimum-variance portfolios in the unconstrained ("full") and sparse ("graph") model. Portfolio weights are computed using the $Var(\mathbf{Y}_t|D_{t-1})$.

For each model at each time point, mean-variance and minimum-variance portfolios were built. Cumulative returns for both investment strategies under all models are displayed in Figures 5.2 and 5.3. In line with the results presented in Chapter 4, in all scenarios considered, the graphical model is more profitable than the full model reinforcing the idea that sparsity can indeed help in portfolio allocation process.

Figure 5.4 displays the optimal *ex ante* risk ratio for the full model relative to the constraint one (mean-variance portfolio). This shows that the estimated portfolio variance is likely to be smaller under the model with graphical structure

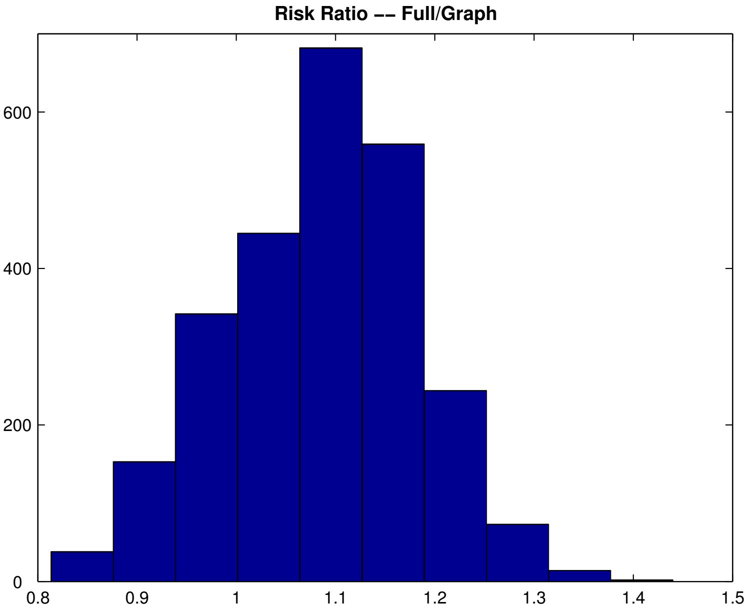


Figure 5.4: *Ex ante* (at time t given D_{t-1}) ratio of standard deviations of optimal portfolios in mean-variance portfolios under the full model relative to the sparse model. Here $\delta = 0.97$.

indicating that the parsimonious model is able to help reduce investment risk. Combined with the fact that more profitable investment opportunities are available, the model with covariance estimation based on a graph provides an *efficient frontier* that dominates, in risk-return sense, the one from the full model.

As previously discussed, one of the reasons behind the good performance of the structured models is possibly the smaller variation of portfolio weights. To explore this notion, a little simulation is performed in the model with $\delta = 0.97$. At each time t samples from the prior of $\mathbf{\Sigma}_t$, in both the full and sparse model, were used as inputs for computing optimal portfolio weights. Figure 5.5 shows the evolution of the implied variance of weights through time. Throughout the entire period of time considered, the optimal amount of money invested in each

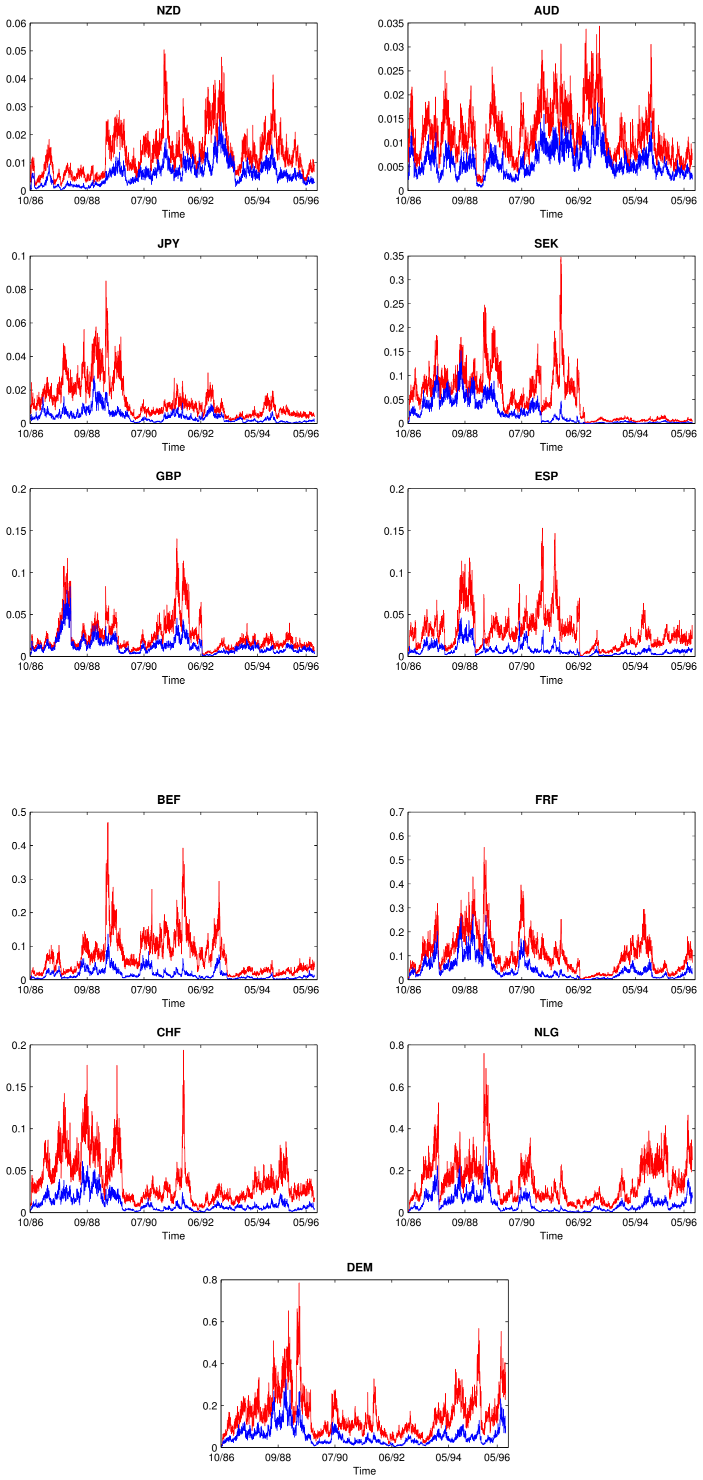


Figure 5.5: Each panel displays the estimated variance of the weights of each asset in the mean-variance portfolio for both the constrained (blue line) and sparse (red line) model.

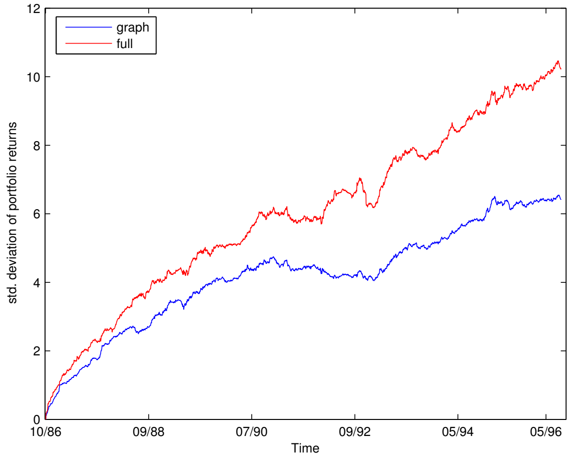
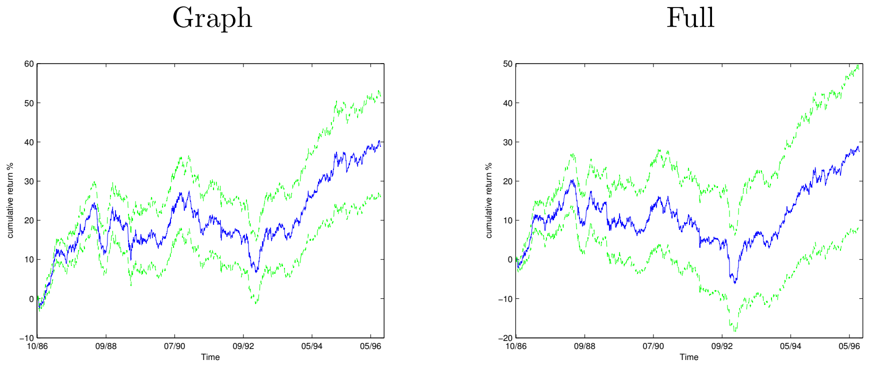


Figure 5.6: Standard deviation of the cumulative returns for the mean-variance portfolio in the full (red line) and sparse (blue line) model.

asset varies considerably more under the full model. The conclusion that there is more uncertainty about the optimal decision in the full model is clear. This fact is highlighted by the distribution of cumulative returns presented in Figure 5.7 where the sequential interval around the expected cumulative return is much wider for the full model. Figure 5.6 clarifies this perception by comparing the estimated standard deviation of cumulative return in the two models.

5.5.2 Portfolio Allocation in the S&P 500

To further explore the performance portfolios built from DLMs with graphical structure, I now present a large scale example using $p = 346$ securities forming part of the S&P500 stock index. These are all the companies that remain part of the stock index from January 1999 until December 2004 creating a set of $t = 1,508$ observations. Again, the goal of the application is to compare the performance of dynamic portfolios built from models with graphical structure relative to the full


Graph

Full

Figure 5.7: Distribution of cumulative returns for the mean-variance portfolio with $\delta = 0.97$ in the sparse (left) and full (right) model.

model. This problem illustrates one focal point of research in portfolio allocation theory where the development of models to efficiently deal with large set of assets is of great interest.

In order to determine the graphs to be used in the example I performed a Metropolis search (restricted to decomposable graphs) using the first 1,200 observations in the dataset. It is important to point out that the only required modification to the search method described in Chapter 2 is in computing the marginal likelihood. In the dynamic set up, the marginal likelihood of the data given G is computed by

$$p(\mathbf{Y}_{1:T}|G) = p(\mathbf{Y}_T|D_{T-1}, G)p(\mathbf{Y}_{T-1}|D_{T-2}, G) \dots p(\mathbf{Y}_1|D_0, G), \tag{5.20}$$

where for 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 5.2. In the search, the prior edge inclusion probability was set to 0.5 so graphs of multiple dimensions could be explored. The initial 1,200 observations were used to generate the prior distribution representing D_0 and the final 308 observations were sequentially modeled as in (5.18). In all models a discount factor of 0.98 was used.

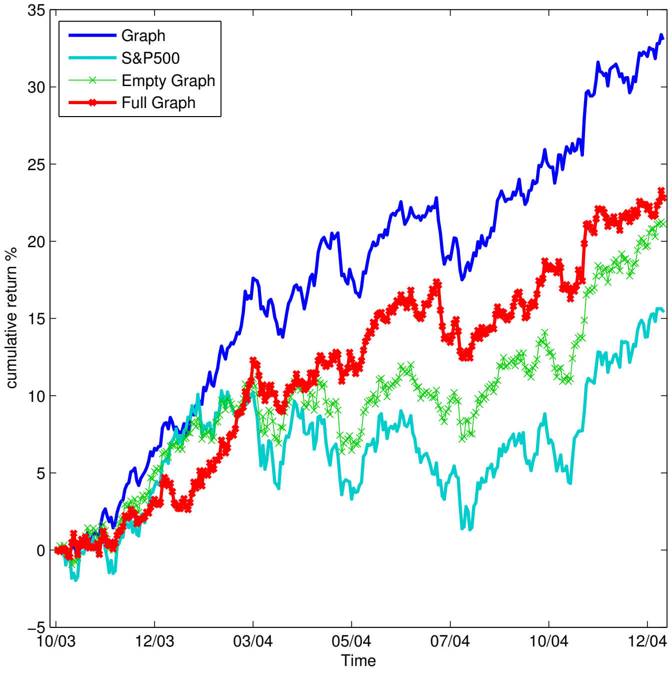


Figure 5.8: S&P 500 portfolios: Cumulative returns for portfolios built from the top graph, the full graph and the empty graph. Cumulative returns of the S&P500 provide a benchmark for the portfolios created in the example.

Figure 5.8 displays the cumulative returns for the "top" graph (29,181 edges) compared to the full graph (59,685 edges) and the empty graph (no edges). As a benchmark the actual returns for the S&P500 are also presented. Again it is clear that the imposition of constraints in the covariance matrix generate more profitable investment opportunities.

The point made earlier about variation in the portfolio weights is highlighted by Figure 5.9 where for four randomly selected companies, the portfolio weights under the graphical model show a more stable behavior through time, in line with the results of the exchange rate example.

One very interesting result of this analysis appears in Figure 5.10 where the

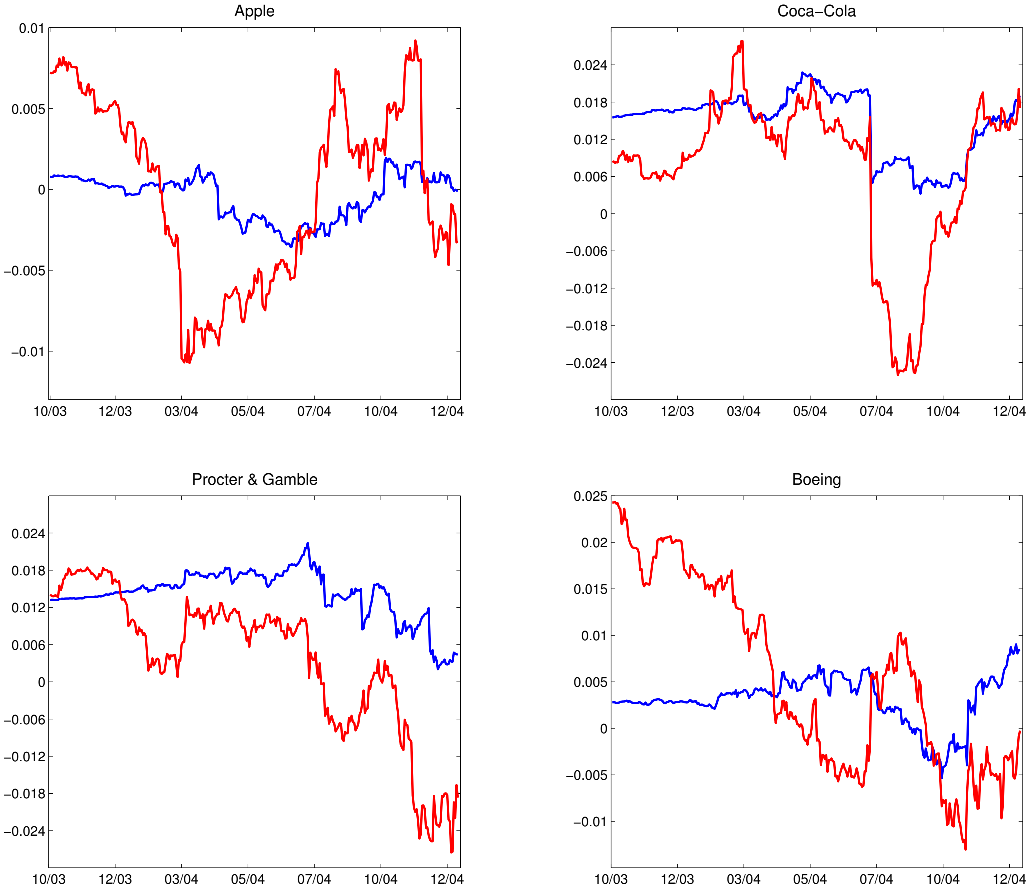


Figure 5.9: Evolution of portfolio weights for 4 companies in both the full graph (red line) and top graph (blue line).

cumulative returns for 5 different graphs are presented. These graphs are a subset of the top 100,000 graphs visited, representing the 25%, 50% and 75% percentile, along with the top graph and the smallest graph in the top group. The fact that the cumulative returns are very similar for all 5 graphs indicates that for the purpose of investment opportunities, a small number of edges are really relevant. Exploring subgroups of graphs that generate "equivalent" cumulative returns may generate a lot of insights about portfolio allocation theory in connection with

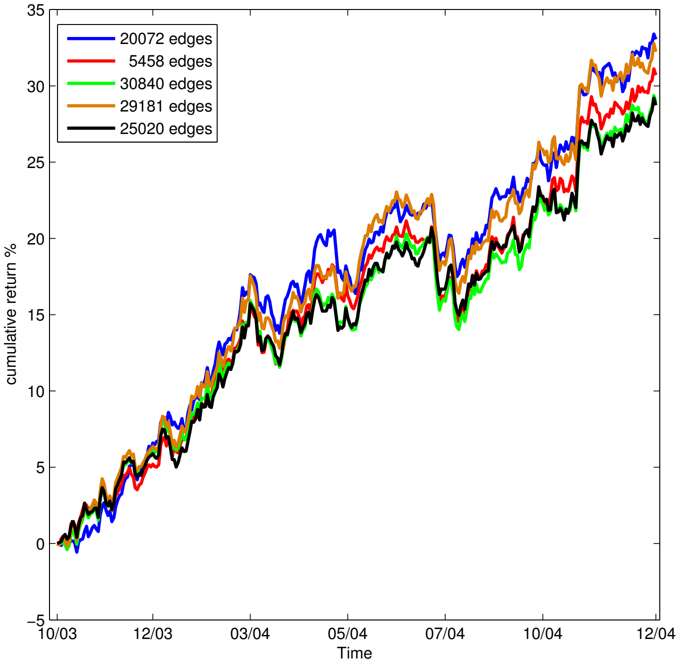


Figure 5.10: S&P 500 portfolios: These represent cumulative returns for portfolios built from 5 different graphs in the top 100,000 graphs visited. The top graph has 29,181 edges while the smallest graph in the top 100,000 has 5,458 edges. The graph with 20,072 edges represents the 25% percentile, the one with 25,020 edges is the median and the one with 30,840 edges is the 75% percentile.

covariance selection models. The development of financial theory is outside the scope of this dissertation but this is definitely an important research road that I intent to explore in the near future.

5.6 Model Uncertainty

In all examples presented, model uncertainty evaluation was performed in a static fashion, i.e, graphs were selected based on a fixed set of observations with Expression (5.20) being the only modification necessary before any model selection strategy described in Chapter 3 is performed.

Each DLM is specified for a given graph G but in general, it is important to recognize that any single DLM (or graph) is inadequate and considering multiple alternatives is fundamental in generating forecasting models that incorporate all sources of uncertainty.

One possible alternative is to work with combinations of DLMs is what is called *multi-process* models as defined in West and Harrison (1997). Any single DLM represents a single process with the combination of several DLMs defining a multi-process model. These combinations are implemented using discrete probability mixtures of DLMs, and so multi-process models can simply be viewed as mixture models. This set up is also very similar to what is called *mixture-of-experts* in the machine learning literature (Carvalho and Tanner, 2005) where each individual forecasting model is one "expert".

West and Harrison (1997) describes in detail the theory of multi-process and established two distinct alternatives: multi-process **class I** and **class II**.

Let $\mathcal{G}$ denote a set of graphs so that for each $G \in \mathcal{G}$, $M_t(G)$ represents the DLM specified by G at time t. If for some $G_0 \in \mathcal{G}$, $M_t(G_0)$ holds for all t, a single DLM is viewed as appropriate for all time but the uncertainty about what is the "true" graph G remains. This framework is defined as **class I** and given a set $\mathcal{G}$, each corresponding DLM is analyzed as usual, producing a sequence of prior, posteriors and forecast distributions that when weighted by $p(G|D_t)$, the posterior probability of graph $G \in \mathcal{G}$ at time t, generate the desired mixture model. Our ability to sequentially calculate $p(G|D_t)$, in closed form, makes this alternative a very attractive one as multiple DLM can efficiently be implemented in parallel, preserving the sequential, on-line nature of these models. The probabilities

$p(G|D_t)$ will change over time as different graphs may best describe the data series at different time intervals.

Multi-process **Class II** models are defined in a way that for each time t, the graph G takes a value in the set $\mathcal{G}_t$, with $\mathcal{G}_t$ varying over time. This represents a much more realistic modeling view but its implementation is very challenging due to the dimensionality of the space of graphs. Selecting a different set of graphs at each time point can be computationally intractable.

The development of sequential model selection procedures that address uncertainty about graphs while allowing for efficient on-line updates is an open research area and one of key importance in further applications of DLMs in real forecasting problems.

5.7 Discussion

By combining dynamic linear models with decomposable graphical models this chapter defines a new class of DLMs that incorporates conditional independence structure in the cross-sectional precision matrix of a set of time series. The use of the hyper-inverse Wishart provides a conjugate framework that allow for sequential updating and on-line predictions.

Models with time varying covariance structure were also described where the extensions of results related to the Bartlett decomposition of HIW matrices in conjunction with discount ideas provides a formal justification for the sequential update of $\mathbf{\Sigma}_t$.

As shown in the portfolios examples, sparsity in modeling covariance structure of asset returns has a huge impact in reducing investment uncertainty. The applications described are simple but able to touch on fundamental problems in high-dimensional asset allocation that hopefully will help the development of both theoretical and empirical results relating conditional independence structure and optimal portfolios.

Chapter 6

High-Dimensional Sparse Factor Models and Evolutionary Model Search

Latent factor analysis consists of models that attempt to explain the variation of a large set of random variables in terms of a small number of unobservable factors. These models explain patterns of association among variables by identifying sources of variation that are common to groups of variables and separate these from idiosyncratic, variable-specific noise. As a model for covariance matrices in multivariate normal problems, factor models represent an alternative to Gaussian graphical models where lower dimensional structure comes directly in the covariance matrix and not its inverse. Factor models are not only another parsimonious option to modeling covariance matrices, but also a way to interpret and understand complex relationships between large number of variables through the analysis of a smaller set of common variational components.

Initial work on factor analysis dates back to the beginning of last century with the study of human abilities by Spearman (1904) followed by a push in the 30's and 40's within the field of psychology. In statistics, Lawley and Maxwell

(1971) and Bartholomew (1984) are two fundamental references with important developments in estimation and testing procedures. Press (1982) is another basic reference with important discussions about identifiability of factor models. Martin and McDonald (1975) and Press and Shigemasu (1989) are early examples of Bayesian inference in factor models, but it is not until MCMC simulation methods become available that more developments are made in works by Geweke and Zhou (1996), Arminger and Múthen (1998), Aguilar and West (2000) and Lopes and West (2004).

Motivated by high-dimensional gene expression problems, West (2003) introduce sparse latent factor models where the dependencies among very many variables are explained by factors that typically relate to a small number of variables, represented by a factor loadings matrix with many zeros.

In this chapter, I describe latent factor models for multivariate analysis in very high dimensions where the use of sparsity inducing priors establishes parsimonious relationships between high-dimensional variables and underlying lower-dimensional latent factors. A key methodological development involves a novel evolutionary stochastic search that addresses important issues of model specification; the approach defines and fits models for high-dimensional problems through an evolutionary process that gradually expands the dimension of the sample space and the dimension of the latent components. It is often the case that we are interested in defining models on subsets of scientific interest and than enriching that analysis by including variables that appear to be related. This is the case in the genomics studies motivating this work where the methodology helps define, enrich and characterize relevant pathways of biological activity by the decomposition of relationships in measured mRNA levels on a genome-wide scale.

6.1 Sparse Factor Models

Let $\mathbf{x}$ be a p-dimensional zero-mean, normal distributed random vector with covariance matrix given by $\mathbf{\Sigma}$. In a basic k-dimensional latent factor model, the i^{th} observation of $\mathbf{x}$ is modeled as

$$\mathbf{x}_i = \mathbf{B}\mathbf{f}_i + \boldsymbol{\nu}_i \tag{6.1}$$

with the following components:

- $\mathbf{B}$ is a $p \times k$ matrix of unknown factor loadings, the factor **loadings matrix**, with elements $\beta_{g,j}$ for $g = 1, \dots, p$ and $j = 1, \dots, k$. Write $\boldsymbol{\beta}'_g$ and $\mathbf{b}_j$ as the vectors representing each row and column of $\mathbf{B}$ respectively.
- $\mathbf{f}_i$ is k-vector of latent **factor scores** with standard normal prior $\mathbf{f}_i \sim N(\mathbf{0}, \mathbf{I})$.
- $\boldsymbol{\nu}_i$ is a p-dimensional vector of independent, idiosyncratic noise, with $\boldsymbol{\nu}_i \sim N(\mathbf{0}, \mathbf{\Psi})$ where $\mathbf{\Psi} = diag(\psi_1, \dots, \psi_p)$.

Further, it is assumed that the set of latent factors and noise terms are independent and mutually independent, i.e., $\mathbf{f}_i \perp\!\!\!\perp \boldsymbol{\nu}_l$, $\mathbf{f}_i \perp\!\!\!\perp \boldsymbol{\nu}_i$, $\mathbf{f}_i \perp\!\!\!\perp \boldsymbol{\nu}_l$ and $\boldsymbol{\nu}_i \perp\!\!\!\perp \boldsymbol{\nu}_l$ for all i, l, $(i \neq l)$. So, for each g the element $x_{g,i}$ of $\mathbf{x}_i$ is given by $x_{g,i} = \boldsymbol{\beta}'_g\mathbf{f}_i + \nu_{g,i}$, where $\nu_{g,i} \sim N(0, \psi_g)$. From these assumptions the covariance structure of $\mathbf{x}$ is constrained by the factor decomposition and takes the form

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{B}' + \mathbf{\Psi}, \tag{6.2}$$

explicitly separating the commonalities ($\mathbf{B}\mathbf{B}'$) from specificities ($\mathbf{\Psi}$) in the variation of $\mathbf{x}$.

6.1.1 Identification

In order to be mathematically identifiable, the k-factor model must be further constrained so that the decomposition in (6.2) has an unique solution. A detailed discussion of identification issues appears in Aguilar (1998) and can be summarized as:

(i) If $\mathbf{B}$ is not full rank the model is not identifiable. If rank($\mathbf{B}$) $= t < k$ it is possible to write $\mathbf{\Sigma} = \mathbf{B}^*\mathbf{B}^{*\prime} + \mathbf{\Psi}$ where $\mathbf{B}^* = \mathbf{B} + \mathbf{M}\mathbf{Q}'$, where $\mathbf{M}$ and $\mathbf{Q}$ are any two matrices of size $p \times (k-t)$ and $k \times (k-t)$ such that $\mathbf{B}\mathbf{Q} = \mathbf{0}$, $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$, and $\mathbf{M}\mathbf{M}'$ is diagonal.

(ii) Orthogonal rotation of factors is another problem. For any orthogonal matrix $\mathbf{T}$, the model can be re-written as a function of $\mathbf{B}^* = \mathbf{B}\mathbf{T}'$ and $\mathbf{f}_i^* = \mathbf{T}\mathbf{f}_i$.

(iii) To guarantee a unique solution for the system of equations generated by (6.2), the number of parameters in the model cannot exceed the total number of parameters in $\mathbf{\Sigma}$. This leads to the constraint that $p(p-1)/2 \geq kp + p$ which establishes a natural upper bound for k.

There are many ways one can impose constraints on $\mathbf{B}$ to address the problems above. Traditional solutions include forcing $\mathbf{B}$ to be orthogonal and constraining $\mathbf{B}'\mathbf{\Psi}\mathbf{B}$ to be diagonal. Throughout this chapter, however, the solution developed by Geweke and Zhou (1996) is preferred, where the loadings matrix takes the form,

$$\mathbf{B} = \begin{pmatrix} \beta_{1,1} & 0 & 0 & \cdots & 0 & 0 \\ \beta_{2,1} & \beta_{2,2} & 0 & \cdots & 0 & 0 \\ \beta_{3,1} & \beta_{3,2} & \beta_{3,3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \beta_{k-1,1} & \beta_{k-1,2} & \beta_{k-1,3} & \cdots & \beta_{k-1,k-1} & 0 \\ \beta_{k,1} & \beta_{k,2} & \beta_{k,3} & \cdots & \beta_{k,k-1} & \beta_{k,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \beta_{p,1} & \beta_{p,2} & \beta_{p,3} & \cdots & \beta_{p,k-1} & \beta_{p,k} \end{pmatrix} \tag{6.3}$$

where $\beta_{i,i} > 0$ for $i = 1, \ldots, k$. This way, $\mathbf{B}$ is full rank and is the unique matrix such that $\mathbf{B}\mathbf{B}'$ is positive definite (Theorem A9.8 Muirhead, 1982). This solution gives weight to the k lead variables in determining the factors, making the order of the first k variables a key modeling decision with direct impact in fit and interpretation of the model. We refer to the first k variables, in the order listed, as the founders of the factors and its choice is one of the problems addressed later on by evolutionary model search (Section 6.4).

6.2 Sparsity Priors

In models for many variables, sparsity modeling aims to induce many zeros in variable-factor relationships - i.e., $\mathbf{B}$ is sparse. The pattern of non-zero values is unknown and to be estimated. A *priori*, each $\beta_{g,j}$ may be zero or take some non-zero value, so that relevant priors should mix point masses at zero with distributions over non-zero values as in standard Bayesian "variable selection" analysis in regression and other areas (Clyde and George, 2004; George and McCulloch, 1993; Raftery *et al.*, 1997). The usual sparsity prior has the form

$$\beta_{g,j} \sim (1 - \pi_j)\delta_0(\beta_{g,j}) + \pi_j N(0, \tau_j), \tag{6.4}$$

independent over g, with $\delta(\cdot)$ being a Dirac delta function at zero. In this set-up there is a common chance π_j of non-zero loadings on factor j for all variables; this base-rate of inclusion on each factor is estimated under a prior that favors small values. A modification is required for the diagonal elements of $\mathbf{B}$ since the identifiability constraint requires that each $\beta_{g,g} > 0$; thus the normal component of (6.4) is adapted to $N(0, \tau_j)I(\beta_{g,g} > 0)$, for $g = 1, \ldots, k$, where $I(\cdot)$ is an indicator function. West (2003) initiated this idea in factor models. However, the use of these standard priors suffers from a critical practical problem that is exacerbated as the number of variables p increases; this is a generic issue impacting on the use of these point-mass/mixture priors in model selection and all other areas. The problem is that with very large p the posterior probabilities for $\beta_{g,j} \neq 0$ tend to be quite spread out over the unit interval leading to an unintuitive high level of uncertainty concerning whether or not $\beta_{g,j} = 0$, for a non-trivial fraction of variables. A solution for this problem arises with the adaptation of the ideas developed by Lucas *et al.* (2006), where a hierarchical prior for $\beta_{g,j}$ reflects the viewpoint that for many variables the probability of association with any one factor is zero (or very small) and for a small set of variables it will be high. The prior takes the form

$$\beta_{g,j} \sim (1 - \pi_{g,j})\delta_0(\beta_{g,j}) + \pi_{g,j} N(0, \tau_j) \tag{6.5}$$

$$\pi_{g,j} \sim (1 - \rho_j)\delta_0(\pi_{g,j}) + \rho_j Be(a_j m_j, a_j(1 - m_j)). \tag{6.6}$$

where $Be(am, a(1-m))$ is a beta distribution with mean m and precision parameter $a > 0$. Each ρ_j has a prior that quite heavily favors very small values, such as $Be(sr, s(1 - r))$ where $s > 0$ is large and r is a very small prior probability of non-zero values. Note that, on integrating out the variable-specific probabilities $\pi_{g,j}$ from the prior for $\beta_{g,j}$ in Equation (6.5), we obtain a similar distribution to (6.4), with $\pi_{g,j}$ simply replaced by $E(\pi_{g,j}|\rho_j) = \rho_j m_j$; this is precisely the traditional variable selection prior discussed above, with the common base-rate of non-zero factor loadings set at $\rho_j m_j$. The insertion of the additional layer of uncertainty between the base-rate and the new $\pi_{g,j}$ reflects the view that many (as represented by small values of ρ_j) of the loadings will be zero for sure, and permits the separation of significant factor loadings from the rest. Now the model has the ability to much more effectively detect non-zero loadings, and to induce very substantial shrinkage towards zero for many, many elements of $\mathbf{B}$ – effectively extracting signal and resolving the implicit multiple comparison problem through an appropriately structured hierarchical model.

6.3 Model Completion and MCMC Implementation

To complete the model, specification of priors for the idiosyncratic variance components in $\mathbf{\Psi}$ and for all τ_j $(j = 1, \ldots, k)$ are required. The elements of $\mathbf{\Psi}$ are assumed to be independent with a rather diffuse (proper) common inverse gamma prior, $\psi_g^{-1} \sim Ga(\frac{a_\psi}{2}, \frac{b_\psi}{2})$, for $g = 1, \ldots, p$. Similarly, we assume conditional independent priors for the variances of the Normal component of the prior for $\beta_{g,j}$, where $\tau_j^{-1} \sim Ga(\frac{a_\tau}{2}, \frac{b_\tau}{2})$ for all $j = 1, \ldots, k$.

For a specified k and a given order of the initial k variables, estimation of the model is done by posterior simulation via MCMC, which can be effectively implemented in a Gibbs sampler format. To establish notation, for any quantity Λ - a subset of the full set of parameters, latent factors and variables - let $p(\Lambda|-)$ represent the complete conditional posterior distribution of Λ given $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and all other parameters. The set of full conditional posteriors is now described.

Sampling p(*$\mathbf{f}|-$*)

Let $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)$. Conditional on $\mathbf{B}$ and $\mathbf{\Psi}$ and based on the conditional independence assumptions of the model, the posterior distribution of $\mathbf{F}$ takes the form

$$\begin{aligned} p(\mathbf{F}|\mathbf{X}, \mathbf{B}, \mathbf{\Psi}) &\propto p(\mathbf{X}|\mathbf{F}, \mathbf{B}, \mathbf{\Psi})p(\mathbf{F}|\mathbf{\Psi}, \mathbf{B}) \\ &= \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{f}_i, \mathbf{B}, \mathbf{\Psi}) \prod_{j=1}^{n} p(\mathbf{f}_j) \\ &= \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{f}_i, \mathbf{B}, \mathbf{\Psi})p(\mathbf{f}_i) \end{aligned} \tag{6.7}$$

which shows that the update of each $\mathbf{f}_i$ can be carried out independently. Given $\mathbf{f}_i \sim N(\mathbf{0}, \mathbf{I})$ and that marginally $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{BB}' + \mathbf{\Psi})$ the joint distribution of $\mathbf{f}_i$ and $\mathbf{x}_i$ is a multivariate normal

$$\begin{pmatrix} \mathbf{x}_i \\ \mathbf{f}_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{BB}' + \mathbf{\Psi} & \mathbf{B} \\ \mathbf{B}' & \mathbf{I} \end{pmatrix}\right]. \tag{6.8}$$

Using standard results of multivariate normals, the posterior conditional distribution of $\mathbf{f}_i$ is simply given by

$$(\mathbf{f}_i|\mathbf{x}_i) \sim N(\mathbf{B}'[\mathbf{BB}' + \mathbf{\Psi}]^{-1}\mathbf{x}_i, \mathbf{I} - \mathbf{B}'[\mathbf{BB}' + \mathbf{\Psi}]^{-1}\mathbf{B}). \tag{6.9}$$

Sampling p(*$\boldsymbol{\beta}, \boldsymbol{\pi}|-$*)

In order to improve mixing, the update of elements $\beta_{g,j}$ of $\mathbf{B}$ and their sparsity-governing probabilities $\pi_{g,j}$ is performed in blocks. We sample the bivariate full conditional distribution for each pair $\{\beta_{g,j}, \pi_{g,j}\}$ by composition, starting from the conditional marginal $p(\beta_{g,j}|-)$ followed by $p(\pi_{g,j}|\beta_{g,j}, -)$. The conditional independence of $x_{g,i}$ given $\mathbf{f}_i$, for all $g = 1, \dots, p$, allows us to sample, for a fixed factor j, each of the pairs $\{\beta_{g,j}, \pi_{g,j}\}$ independently. For factor j, let $x^*_{g,i} = x_{g,i} - \sum_{l=1,l\neq j}^{k} \beta_{g,l} f_{l,i}$, so that $x^*_{g,i} \sim N(\beta_{g,j} f_{j,i}, \psi_g)$ and, consider only the unconstrained elements of $\mathbf{B}$ where $g \neq j$. We obtain the conditional marginal for $\beta_{g,j}$ integrating out $\pi_{g,j}$ from the prior, i.e.,

$$p(\beta_{g,j}) = (1 - \rho_j m_j)\delta_0(\beta_{g,j}) + \rho_j m_j N(0, \tau_j). \tag{6.10}$$

This gives the following posterior update:

$$\begin{aligned} p(\beta_{g,j}|-) &\propto \prod_{i=1}^{n} p(x^*_{g,i}|\beta_{g,j} f_{j,i}, \psi_g) p(\beta_{g,j}) \\ &= \prod_{i=1}^{n} N(x^*_{g,i}|\beta_{g,j} f_{j,i}, \psi_g) \left[(1 - \rho_j m_j)\delta_0(\beta_{g,j}) + \rho_j m_j N(0, \tau_j)\right] \\ &= (1 - \hat{\pi}_{g,j})\delta_0(\beta_{g,j}) + \hat{\pi}_{g,j} N(\mu_{g,j}, C_{g,j}) \qquad (6.11) \\ & \qquad (6.12) \end{aligned}$$

where $C_{g,j} = \left(\frac{\sum_{i=1}^{n} f_{j,i}^2}{\psi_g} + \tau_j^{-1} \right)^{-1}$, $\mu_{g,j} = C_{g,j} \left(\sum_{i=1}^{n} f_{j,i} x^*_{g,i} \right) \psi_g^{-1}$ and $\beta_{g,j} \neq 0$ with odds

$$\left(\frac{\hat{\pi}_{g,j}}{1 - \hat{\pi}_{g,j}} \right) = \frac{N(0|0, \tau_j)}{N(0|\mu_{g,j}, C_{g,j})} \frac{\rho_j m_j}{1 - \rho_j m_j}. \tag{6.13}$$

Second, for the constrained diagonal elements of $\mathbf{B}$, the conditional posterior is simply given by

$$(\beta_{j,j}|-) \sim N(\mu_{j,j}, C_{j,j}) I(\beta_{j,j} > 0) \tag{6.14}$$

with similar formulas for $\mu_{j,j}$ and $C_{j,j}$.

After sampling $\beta_{g,j}$ marginally, we need to simulate $\pi_{g,j}$ from the posterior conditional $p(\pi_{g,j}|\beta_{g,j}, -)$. First, let $\beta_{g,j} = 0$ so we can write the full conditional posterior of $\pi_{g,j}$ as

$$\begin{aligned} p(\pi_{g,j}|\beta_{g,j}=0,-) &\propto (1-\pi_{g,j})\delta_0(\beta_{g,j})\,[(1-\rho_j)\delta_0(\pi_{g,j})+ \\ &\qquad \rho_j Be(\pi_{g,j}|a_j m_j, a_j(1-m_j))\,] \\ &= (1-\rho_j)\delta_0(\pi_{g,j}) + \rho_j(1-m_j)Be(\pi_{g,j}|a_j m_j, a_j(1-m_j)+1) \\ &= (1-\hat{\rho}_j)\delta_0(\pi_{g,j}) + \hat{\rho}_j Be(\pi_{g,j}|a_j m_j, a_j(1-m_j)+1) \end{aligned} \tag{6.15}$$

where

$$\left(\frac{\hat{\rho}_j}{1-\hat{\rho}_j}\right) = \frac{(1-m_j)\rho_j}{1-\rho_j}. \tag{6.16}$$

The distribution in (6.15) states that when $\beta_{g,j} = 0$, $\pi_{g,j}$ is sampled from $Be(\pi_{g,j}|a_j m_j, a_j(1-m_j)+1)$ with probability $\hat{\rho}_j$ and $\pi_{g,j} = 0$ otherwise.

Now, if $\beta_{g,j} \neq 0$,

$$\begin{aligned} p(\pi_{g,j}|\beta_{g,j}\neq 0,-) &\propto \pi_{g,j} N(\beta_{g,j}|0,\tau_j)\,[(1-\rho_j)\delta_0(\pi_{g,j})+ \\ &\qquad \rho_j Be(\pi_{g,j}|a_j m_j, a_j(1-m_j))] \\ &= \pi_{g,j} N(\beta_{g,j}|0,\tau_j)\rho_j Be(\pi_{g,j}|a_j m_j, a_j(1-m_j)) \\ &= \pi_{g,j} Be(\pi_{g,j}|a_j m_j, a_j(1-m_j)) \end{aligned} \tag{6.17}$$

which implies that

$$(\pi_{g,j}|\beta_{g,j}\neq 0,-) \sim Be(a_j m_j + 1, a_j(1-m_j)). \tag{6.18}$$

Sampling $p(\boldsymbol{\tau}|-)$

The full conditional posterior distribution for each τ_j^{-1} can be expressed as

$$\begin{aligned} p(\tau_j^{-1}|-) &\propto \prod_{g=1}^{p} p(\beta_{g,j}|\pi_{g,j},\tau_j)p(\tau_j^{-1}) \\ &= \prod_{g=1}^{p} N(\beta_{g,j}|0,\tau_j^{-1})Ga\left(\tau_j^{-1}\left|\frac{a_\tau}{2},\frac{b_\tau}{2}\right.\right) \qquad (6.19) \end{aligned}$$

and therefore

$$(\tau_j^{-1}|-) \sim Ga\left(\frac{a_\tau + w_j}{2}, \frac{b_\tau + \sum_{g=1}^{p}\beta_{g,j}^2}{2}\right), \qquad (6.20)$$

where $w_j = \sum_{g=1}^{p} \mathbf{1}_{\{\beta_{g,j}\neq 0\}}$ and independently over j.

Sampling $p(\boldsymbol{\Psi}|-)$

Recall that the prior for each diagonal element ψ_g of $\boldsymbol{\Psi}$ is an independent inverse-gamma, $IG(\frac{a_\psi}{2},\frac{b_\psi}{2})$. So, for all $g = 1,\ldots,p$ the full conditional posterior is given by

$$\begin{aligned} p(\psi_g^{-1}|-) &\propto \prod_{i=1}^{n} p(x_{g,i}|\boldsymbol{\beta}_g'\mathbf{f}_i,\psi_g^{-1})p(\psi_g^{-1}) \\ &= \prod_{i=1}^{n} N(x_{g,i}|\boldsymbol{\beta}_g'\mathbf{f}_i,\psi_g)Ga\left(\psi_g^{-1}\left|\frac{a_\psi}{2},\frac{b_\psi}{2}\right.\right) \qquad (6.21) \end{aligned}$$

yielding

$$(\psi_g^{-1}|-) \sim Ga\left(\frac{a_\psi + n}{2}, \frac{b_\psi + \sum_{i=1}^{n}(x_{g,i} - \boldsymbol{\beta}_g'\mathbf{f}_i)^2}{2}\right), \qquad (6.22)$$

independently over g.

***Sampling* $\boldsymbol{p(\rho|-)}$**

Finally, we can sample each ρ_j independently with full conditional posterior given by

$$\begin{aligned} p(\rho_j|-) &\propto \prod_{g=1}^{p} p(\pi_{g,j}|\rho_j)p(\rho_j) \\ &= (1-\rho_j)^{p-j-S_j}\rho_j^{S_j} Be(\rho_j|sr, s(1-r)) \\ p(\rho_j|-) &= Be(sr+S_j, s(1-r)+p-j-S_j) \end{aligned} \tag{6.23}$$

where $S_j = \sum_{g=j}^{p} I(\pi_{g,j} \neq 0)$.

6.4 Evolutionary Model Determination

First some motivating context and discussion. Suppose that genome-wide expression profiles are available in a set of breast cancer tumors and our goal is to understand and explore connections with a particular hormonal pathway – the estrogen receptor (ER) pathway, for example. Thinking of pathways as underlying dimensions of biological activity, sparse factor models arise as a direct way to decompose the associations among genes into components of variation, i.e. factors, representing these pathways. Estimates of the loadings $\mathbf{B}$ and inclusion probabilities $\pi_{g,j}$ help assess the roles played by each of the pathways in the variation of the genes, facilitating interpretation and understanding of biological ties. The problem, however, lies with the fact that fitting a factor model for thousands of variables (in this case up to 30,000) not only requires complex modeling decisions but is also computationally challenging. With the scientific interest being on a specific pathway related to a certain number of genes we don't care at all about many of the variables thus, fitting models for all genes is not an efficient way to aim at this problem. Given a "nucleating" set of genes with known involvement in the target pathway, our goal is to enrich the analysis by identifying other genes that share activity with the underlying components currently defining the instances of biological variation.

With the pathway exploration idea in mind, the approach developed in this section tries to address critical questions of model specification and fit in higher-dimensions by allowing an evolutionary increase in the number of variables entertained by the model generating a more focused analysis of the latent factor structure. Starting from a set of variables with known involvement in one or more pathways of interest, the evolutionary search will sequentially expand the sample space with variables related to the current components of variation while allowing the model to increase its complexity by the inclusion of new factors. The method also serves as a exploratory tool that will, for a given set of variables, help determined the number of factors and the order of the first k variables. All decisions made in each of the steps are based upon MCMC estimates of the sparsity pattern in $\mathbf{B}$ which gives some theoretical justification to the procedure, as it can be viewed as a conditional search within a over-arching proper MCMC.

Next, I describe each of the steps involved in the search.

Ordering the Variables

Given the lower triangular structure of $\mathbf{B}$ the order of the initial variables is key in the estimation of the factors. Ideally, the top variable should be the one heavily associated with the most dominant component of variation while being conditionally independent of the other factors. Variable number two should be very representative of factor two while conditional independent of the subsequent factors, and so on. For a p-dimensional vector $\mathbf{x}$ the following steps aim to create a reasonable order for a k-dimensional factor model:

1. For $T = 1$ to k:

   (i) Fit a model with T factors, without the constraint $\beta_{T,T} > 0$;

   (ii) Compute the posterior mean of $\mathbf{B}$, $\hat{\mathbf{B}}$;

   (iii) Rank the variables by $|\hat{\mathbf{b}}_T|$. Choose the "top" variable to be the T^{th} variable in the list (hereafter, the founder of factor T);

   (iv) Set the constraint $\beta_{T,T} > 0$;

2. Re-fit model with the final order of the initial k variables, subject to **(iv)**.

The sequential inclusion of factors attempt to capture, in order of importance, each of the components of variation and identify the variable that is most related to that factor. This variable should also be the one with the smallest estimated idiosyncratic variation, given the factors included so far, and therefore the ordering will be consistent with the assumptions implied by the upper triangular shape of $\mathbf{B}$. At each step, the positivity constraint of the top variable is removed in order prevent the estimation to be biased towards the variable currently on top which, at that point, is not necessarily a founder. Once the founder is chosen the constraint is put back in, before the model is re-estimated.

Including Variables

Suppose we are dealing with a dataset of size p and that currently a set of $p_{in} < p$ variables is being modeled by a k factor model. The idea is to expand around the current latent components and we are therefore interested in bringing to the model variables that show association with the current factors. To do so, an auxiliary MCMC is implemented and variables are brought into the model based on the posterior estimates of inclusion probabilities and loadings weights. The following describes the procedure:

1. Estimate the model for the p_{in} variables and k factors. Compute $\hat{\mathbf{F}}$.

2. Run an auxiliary MCMC for a k factor model on the $p_{out} = p - p_{in}$ variables not in the model, fixing the values of $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)$ at the posterior estimates $\hat{\mathbf{F}}$ from the k factor model based on the p_{in} current variables;

3. Compute posterior estimates of the inclusion probabilities $\pi_{g,j}$ for all $g = 1, \ldots, p_{out}$ and $j = 1, \ldots, k$.

4. For $g = 1$ to p_{out}, rank the variables by the maximum estimated inclusion probability $\pi_{g,j}$.

5. Include the top z variables in the model;

6. Re-estimate a k factor model for the $p'_{in} = p_{in} + z$ variables.

Running the auxiliary MCMC with fixed values of $\mathbf{F}$ is equivalent of running a Gibbs sampler for variable selection that includes each of the variables outside of the model, and in which the possible regressors are the current estimates of the factor scores. The levels of association with each of the current factors, estimated in $\hat{\mathbf{B}}_{out}$, will tell us where to expand the analysis. Again, the sparsity inducing priors play a central role, providing a direct and formal way to identify the group of candidate variables to be included. The choice of z controls how aggressively the search will expand around the existing, current set of included variables.

Including a Factor

Suppose the current model has p variables and k factors. When the inclusion of a new factor is proposed, an auxiliary MCMC is implemented with $k+1$ factors without the constraint that $\beta_{k+1,k+1} > 0$. The decision to keep the newly proposed factor is based on the estimates of $\pi_{g,k+1}$ for all $g = 1, \ldots, p$. If enough variables are identified with high probability of loading on factor $k+1$, the factor is included. When a factor is included, before refitting the model, the $(k+1)^{th}$ founder has to be selected. Again, the variable with highest $|\hat{\mathbf{b}}_{k+1}|$ will be selected as the $(k+1)^{th}$ founder.

There are two control parameters in this step: (i) the probability cut-off that will determined which variables are loaded in the new factor and (ii) the minimum number of variables loaded in a factor to justify its inclusion. The choice of these parameters are specific to each analysis and will depend on modeling goals.

6.4.1 Evolutionary Search

The combination of the steps described above establishes what we call the evolutionary model determination. Given a set of p-variables as a starting point and starting with $k = k_0$, the search proceeds as follows:

1. Choose the k_0 founders;

2. Fit a model with p variables and $k = k_0$ factors;

3. Try to include up to z variables; Re-fit model;

4. Try to include a new factor; Re-fit model;

5. If no more variables or more factors can be included stop; Else goto step 3.

After choosing the initial founders, we start trying to include variables. Newly included variables may introduce new sources of variation and so we follow with the proposal of a new factor. When a new factor is included a different aspect of covariation is introduced and explored; this may enrich the model by the inclusion of new variables associated with it. We keep iterating between steps 3 and 4 until no more variables or factors are included. This happens when no variables outside of the model meet a pre-determined inclusion probability threshold, or when to few variables are significantly loaded on a newly proposed factor.

Two intermediate steps enhance the performance of the search: (i) Re-selection the founders; (ii) excluding variables. Step (i) is key as the inclusion of new variables could provide better founders for the current factors. As for step (ii), if after fitting the model, some variables are not loaded in any of the existing factors (as estimated in $\hat{\mathbf{B}}$), they are dropped out of the model. This is relevant in helping getting rid of possible "bad" variables that were part of the initial set. As the search goes on, it is generally the case that only variables that are really related to the factors being explored are drawn into the model, and step (ii) will tend to be unnecessary. In more aggressive searches, however, when many variables are included at each step, it is important to keep checking whether all variables in the model are really participating in the activity estimated by the factors. Allowing variables to drop out of the model adds flexibility to the procedure, helping in the exploration and enrichment of the sample space.

The development of the evolutionary search is inspired by the idea that we are able to explore the latent structure of a large vector $\mathbf{x}$ by sequentially understanding the structure of subsets of $\mathbf{x}$. It is possible to think of the evolutionary search as a way to estimate the k factor model for the entire vector $\mathbf{x}$ by running a conditional MCMC where, for the variables and factors out of the current model, the elements of **B** are fixed at zero. Expression (6.24) is a illustration of this where, assuming that the true $k = 6$, at that point 6 variables are in the model with 3 latent factors. As the search evolves, variables and factors are included and fewer values of **B** are forced to zero, as is (6.25).

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \end{pmatrix} = \begin{pmatrix} \beta_{1,1} & 0 & 0 & 0 & 0 & 0 \\ \beta_{2,1} & \beta_{2,2} & 0 & 0 & 0 & 0 \\ \beta_{3,1} & \beta_{3,2} & 0 & 0 & 0 & 0 \\ \beta_{4,1} & \beta_{4,2} & 0 & 0 & 0 & 0 \\ \beta_{5,1} & \beta_{5,2} & 0 & 0 & 0 & 0 \\ \beta_{6,1} & \beta_{6,2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \\ \nu_5 \\ \nu_6 \\ \nu_7 \\ \nu_8 \\ \nu_9 \\ \nu_{10} \end{pmatrix} \tag{6.24}$$

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \end{pmatrix} = \begin{pmatrix} \beta_{1,1} & 0 & 0 & 0 & 0 & 0 \\ \beta_{2,1} & \beta_{2,2} & 0 & 0 & 0 & 0 \\ \beta_{3,1} & \beta_{3,2} & \beta_{3,3} & 0 & 0 & 0 \\ \beta_{4,1} & \beta_{4,2} & \beta_{4,3} & 0 & 0 & 0 \\ \beta_{5,1} & \beta_{5,2} & \beta_{5,3} & 0 & 0 & 0 \\ \beta_{6,1} & \beta_{6,2} & \beta_{6,3} & 0 & 0 & 0 \\ \beta_{7,1} & \beta_{7,2} & \beta_{7,3} & 0 & 0 & 0 \\ \beta_{8,1} & \beta_{8,2} & \beta_{8,3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \\ \nu_5 \\ \nu_6 \\ \nu_7 \\ \nu_8 \\ \nu_9 \\ \nu_{10} \end{pmatrix} \tag{6.25}$$

6.5 Examples

In this section, three examples illustrate the use of sparse factor models and the evolutionary search. I start with a simulated example where the performance of the evolutionary search is tested. A second small example illustrates how the evolutionary search can also be used as a exploratory way to determine k. The

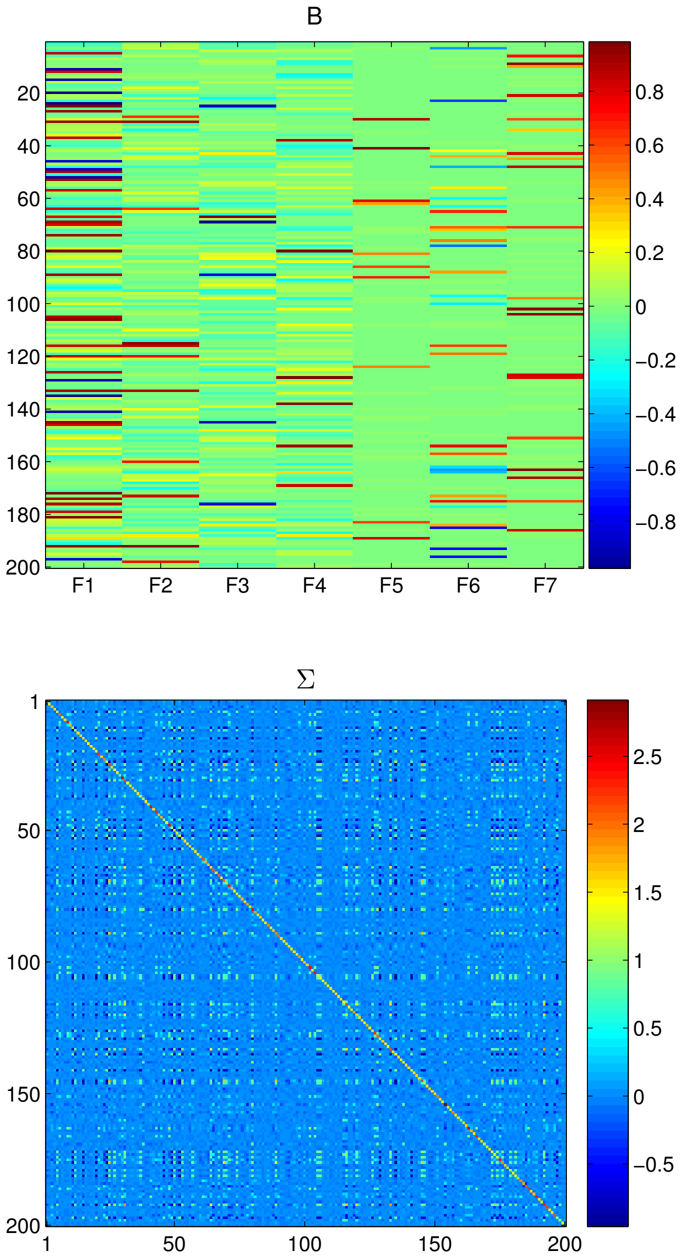


Figure 6.1: Simulated example: The top panel shows the true loadings matrix **B** while the true covariance matrix **Σ** appears in the bottom panel.

Runs	Variables Out
1	80, 128, 138, 163
2	138, 146, 163
3	107, 128, 138
4	80, 128, 146, 163
5	138, 146, 163
6	80, 128, 138, 163
7	80, 138, 163
8	107, 138, 146, 163

Table 6.1: Simulated factor analysis example: “Bad” (8 of 100) runs and relevant variables left out of the model.

section ends with a high-dimensional genomic example where the ER pathway is explored using gene expression data from human breast cancers.

6.5.1 Simulation Study

In order to test the performance of the evolutionary search, a dataset of 200 variables and 100 samples was generated from a 7-factor model where $\mathbf{B}$ is very sparse, with only 93 variables loaded on at least one of the factors. To make the example more realistic, noise was added to the zero elements of $\mathbf{B}$. Figure 6.1 shows images of the true $\mathbf{B}$ and $\mathbf{\Sigma}$. $\mathbf{B}$ was constructed trying to represent some of the problems that motivate this work. Factor 1 represents the main factor, playing a role in the variation of many variables. Factors 2 through 4 try to represent sub-factors highly connected to factor 1 while factors 5 through 7 represent factors with only subtle connections to the main factor of interest. The search was performed 100 times, always starting from a set of 10 variables of which at least 7 were part of the 93 that are actually loaded on the factors; this is consistent with the idea that the goal is to explore around particular “pathways” of interest represented by an initial set of “nucleating” variables. In each run,

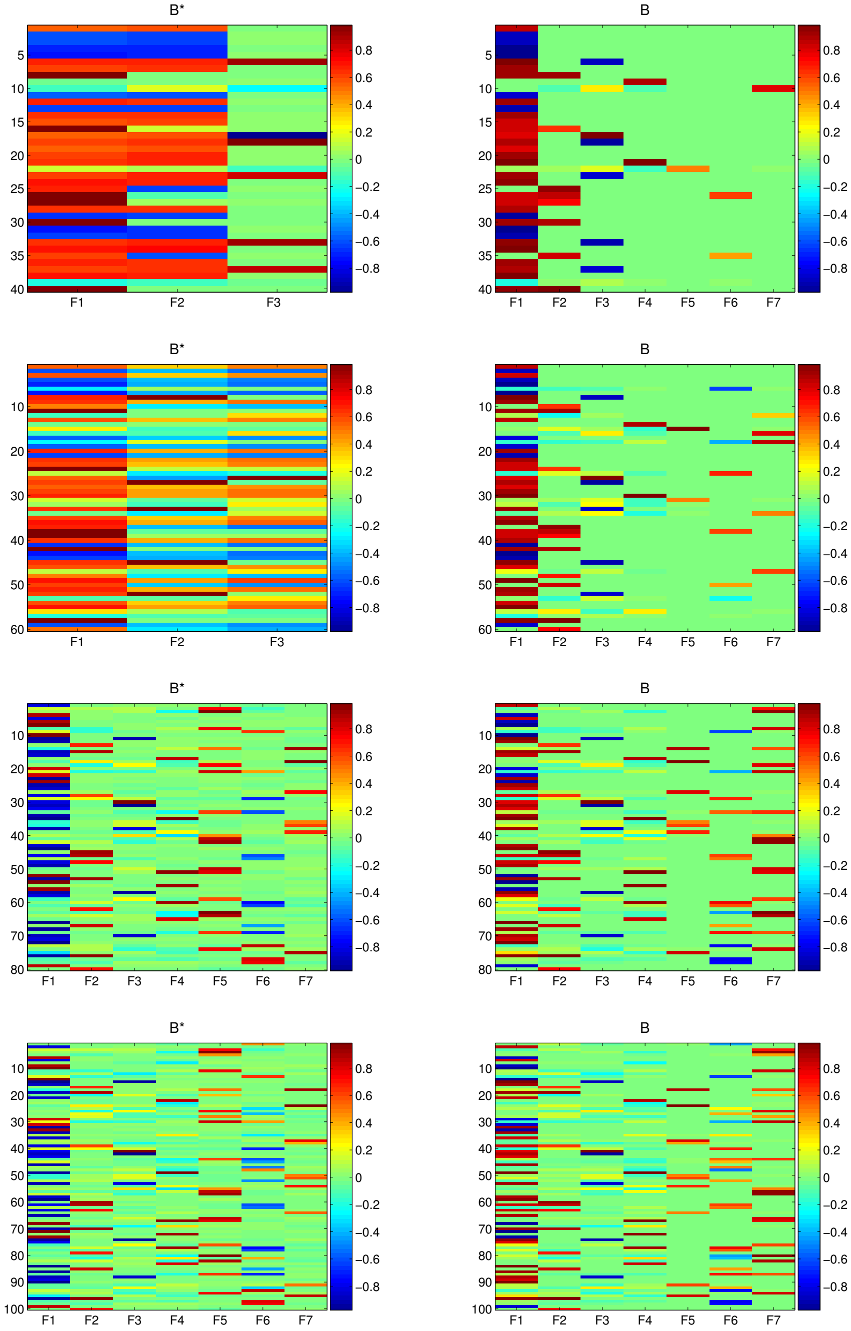


Figure 6.2: Simulated factor analysis example: Snapshots of estimates ($\mathbf{B}^*$) and true values of $\mathbf{B}$ after 40, 60, 80 and 100 variables are included in the model.

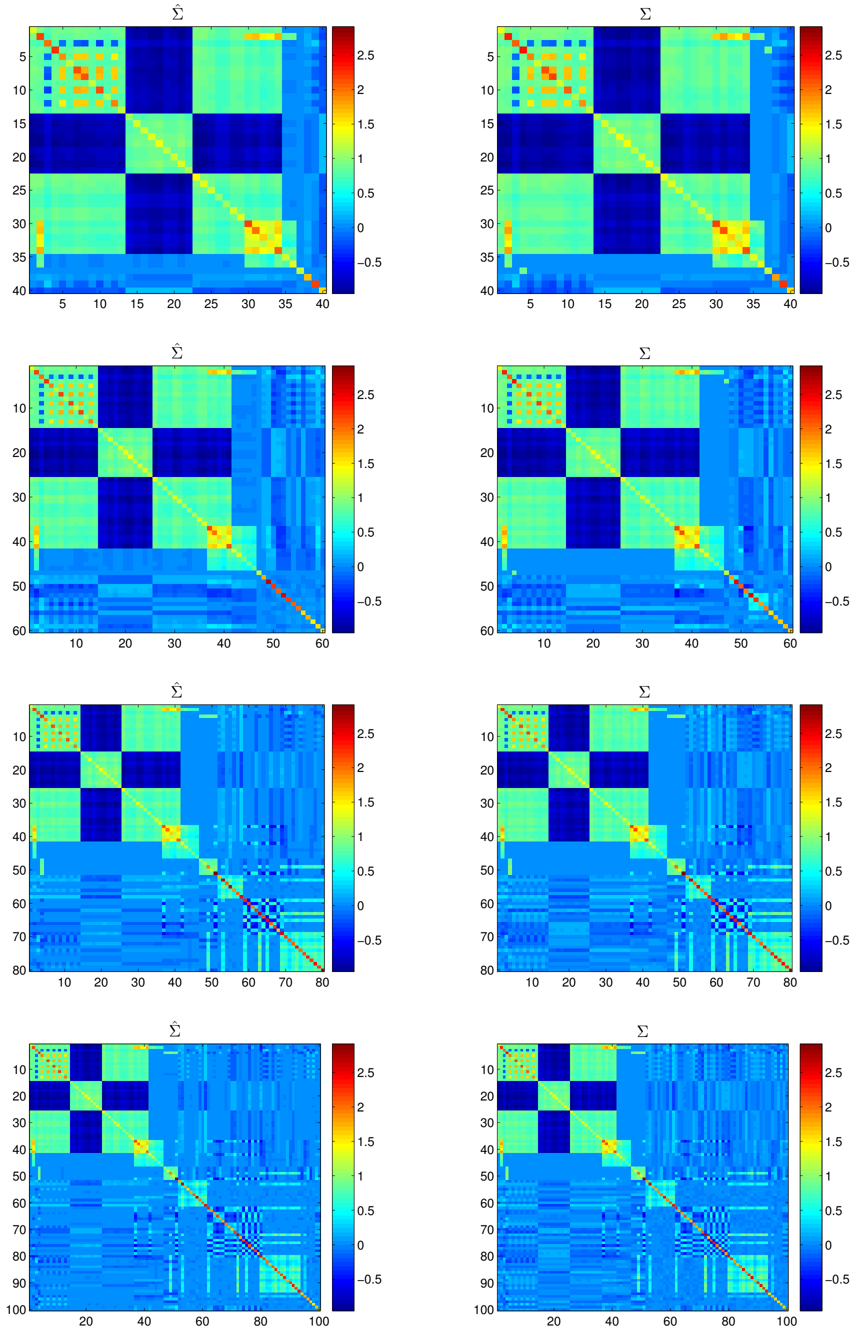


Figure 6.3: Images of estimates ($\hat{\mathbf{\Sigma}}$) and true values of $\mathbf{\Sigma}$ for the variables in the model, in the simulated factor analysis example.. In these images the variables are re-ordered so that all the variables loaded in factor 1 are placed on top, followed by all variables loaded in factor 2 and so on. Again, these snapshots are taken after 40, 60, 80 and 100 variables are included in the model.

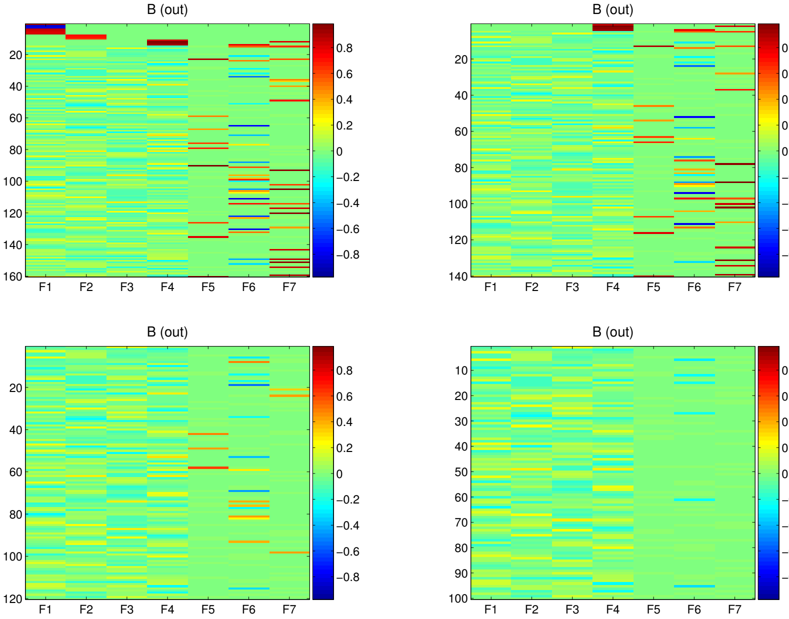
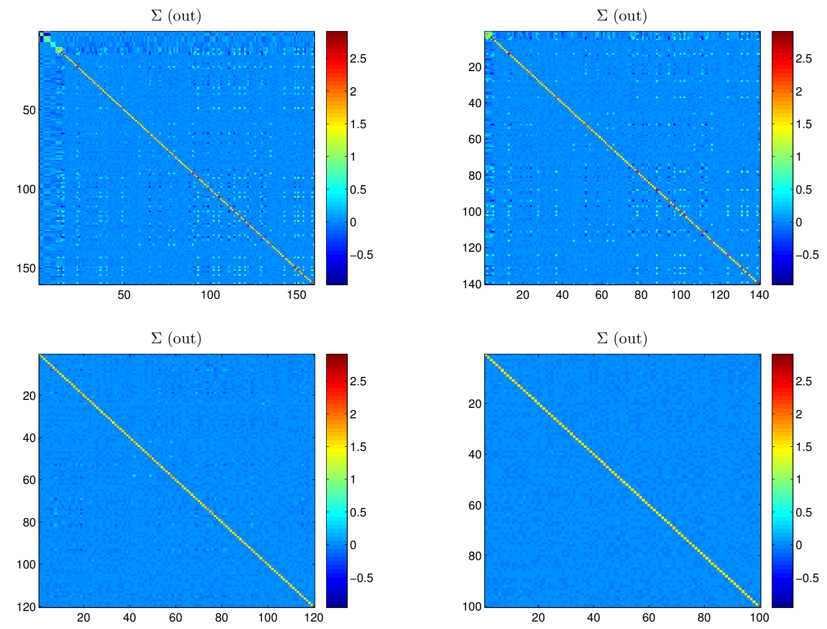


Figure 6.4: Simulated factor analysis example: Images of true loadings **B** of variables out of the model after 40, 60, 80 and 100 variables are included in the model. In the end, only zeros (or very small values) are left out of the model

the search started with one factor and the following parameters/conditions where used:

- Variables were included in the model only if the estimated inclusion probability $\pi_{g,j} > 0.95$ for at least one factor $(j = 1, \ldots, k)$;
- Proposed factor l was included in the model if at least 5 variables showed probability of inclusion $\pi_{g,l} > 0.95$;
- At each step, at most 10 variables were included;
- After 60 variables were in the model, the founders were selected again;
- The search was set to stop after 100 variables were in the model.


Σ (out)

Σ (out)

Σ (out)

Σ (out)

Figure 6.5: Simulated factor analysis example: Images of the true $\mathbf{\Sigma}$ for variables out of the model after 40, 60, 80 and 100 variables are included in the model. The evolutionary search sequentially captures structure in the covariance matrix leaving out of the model only variables the are not related to any of the factors.

In all but 8 of the 100 runs, all 93 "relevant" variables were included and in all runs the final number of factors was 7. Further analysis of the 8 runs that failed to include all relevant variables show that the variables "out" were part of the same small group (Table 6.5.1). Not surprisingly this group represents variables with the lowest percentage of variation explained by the factors (Table 6.5.1) and were left out due to weak association with the seven components of variation.

Figures 6.2, 6.3, 6.4 and 6.5 illustrate the evolution of one run where snapshots of the search were taken with 40, 60, 80 and 100 variables in the model. Figure 6.3 displays estimates of $\mathbf{\Sigma}$ at each step next to the true covariance matrix for the variables in the model. It is clear that the method is able to sequentially capture the structure in $\mathbf{\Sigma}$. This is reinforced in Figure 6.5 where we can see that only

Variable	% of Variation
x_{80}	11.08%
x_{107}	13.37%
x_{128}	12.37%
x_{138}	10.95%
x_{146}	12.29%
x_{163}	10.82%

Table 6.2: Simulated factor analysis example: Percentage of variation explained by all factors for the relevant variables left out of the model in at least one of the "bad" runs.

uncorrelated variables are left out of the model. Figure 6.2 shows how well the estimation of $\mathbf{B}$ is carried out. In the end, the search is able to determine the correct number of factors and identify the correct variables loaded on each of the factors. Note that, in this particular run, factor 7 was the fifth factor included while factor 5 was included last. Also, there a sign change in factors 1 and 6 which relates to the choice of the founders and identification constraint imposed for the top variable. As it can be seen in the estimates of $\mathbf{\Sigma}$ these differences have no impact in estimation or interpretation of the model.

6.5.2 Selecting k

Generally, in factor analysis, the number of factors k is a modeling choice and in most applied work k is used as a control parameter to test sensitivity of predictions and change in interpretation. One of few fully Bayesian attempts to formally make inferences about k appears in Lopes and West (2004) where a reversible jump MCMC is proposed to move around the space of models, avoiding the problem of estimating marginal likelihoods. Their method, however, requires parallel Gibbs samplers for all models considered in order to generate suitable empirical proposals that are used in the RJMCMC. In problems of very high-dimensions where

Currency	Factor 1	Factor 2	Factor 3	Noise
DEM	95.48	0.00	0.00	4.51
AUD	0.99	56.18	0.00	42.81
SEK	58.55	0.00	6.27	35.17
ESP	64.49	0.00	6.28	29.22
GBP	58.34	2.30	3.80	35.54
JPY	41.09	0.00	0.00	58.90
BEF	85.78	0.00	2.06	12.15
FRF	86.85	0.00	2.14	10.99
NZD	3.05	39.20	0.00	57.73
NLG	95.65	0.00	0.00	4.34
CHF	86.10	0.00	0.00	13.89

Table 6.3: Percentage of variation explained by each factor and idiosyncratic noise component in the exchange rate financial example.

the number of factors is really uncertain, this approach becomes computationally very unattractive and infeasible. Again, the evolutionary search provides an exploratory way to determine k which can at least serve as a way to narrow down the number of possibilities for a formal RJMCMC. Given a fixed set of variables of size p the evolutionary search is a forward selection procedure that sequentially expands the dimension of the model based on the sparsity pattern of $\mathbf{B}$. This is very similar to projection pursuit methods (Friedman and Tukey, 1974; Tukey and Tukey, 1981) that iteratively search and remove structure from high-dimensional multivariate datasets by projecting the data into lower-dimensional spaces.

Building on a example that appears in Lopes and West (2004) I apply the evolutionary search to set of returns on international currencies trying to determine k. The data is that presented in Chapters 4 and 5 (Figure 5.1). Starting with $k = 1$, factors were sequentially proposed and accepted when at least 2 variables presented estimated inclusion probability $\pi_{g,j_{new}} > 0.95$. The search stopped with $k = 3$ which is consistent with the information criteria displayed in Table 6.5.2

	$k = 1$	$k = 2$	$k = 3$	$k = 4$
AIC	-4,642	-6,154	**-6,255**	-6,187
BIC	-4,636	-6,142	**-6,237**	-6,152

Table 6.4: AIC and BIC for models with different number of factors (k) in the exchange rate financial example. Both criteria agree with the end point of the evolutionary search.

and similar to the results in Lopes and West (2004). Interpretation of the model with $k = 3$ (Table 6.5.2) is straightforward and also consistent with the graphical model example of Chapter 4. Basically, the first factor (Deutsch Mark as founder) represents the main activity in the EU area, and explains most of the variation in European currencies. Factor 2 identifies the Oceania component (Australian dollar is the founder), the main source of variation for AUS and NZD. The third factor separates out a more subtle effect in some European currencies, heavily influenced by countries outside of the monetary union (GBP and SEK). The Japanese Yen has some of its variation explained by the EU factor but most of it remains in the idiosyncratic noise which is an indication that the model is missing an Asian component- no other Asian country is considered in the model. The same type of conclusion can be drawn from the top graph in Figure 4.1 where the pattern of conditional independencies is consistent with the factors estimated here. In fact, if we construct a graph by truncating small values in $\hat{\mathbf{\Sigma}}^{-1}$ (Figure 6.6) we can identify features similar to those in Figure 4.1.

6.5.3 Exploring the ER Pathway in Breast Cancer Genomics

I now present an analysis of gene expression profiles from DNA microarrays assays of mRNA from breast tumors. The data consists of 171 samples from tumors

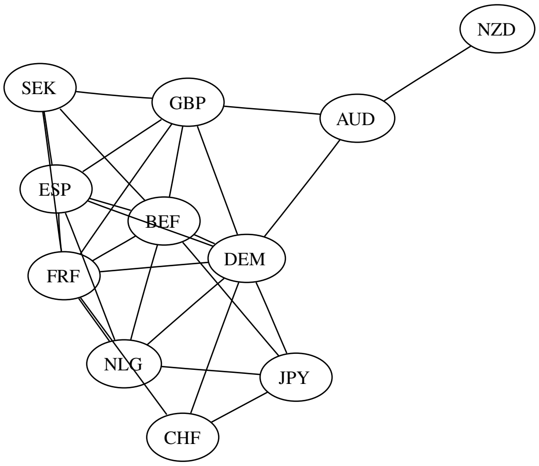


Figure 6.6: Conditional independence graph implied by $k = 3$ factor model in the exchange rate financial example.

where expression profiles of about 12,000 genes were taken in Affymetrix Human U95Av2 GeneChips and processed using the current standard RMA (Irizarry *et al.*, 2003b,a) to generate summary estimates of expression levels of each gene in each sample. The data is then transformed to log_2 expression values and (an illustration of data is shown in Figure 6.7) after screening the dataset for genes showing limited variation over samples, 5,000 genes were considered. This data comes from the Sun-Yat Sen Cancer Center in Taipei and it has been thoroughly analyzed in studies by West *et al.* (2001), Huang *et al.* (2003) and Pittman *et al.* (2004) where the central goal was the identification of aggregate patterns of gene expression capable to predict lymph node metastasis and cancer recurrence. The goal of the analysis is to explore hormonal and growth pathways and their interactions, with special attention to the estrogen receptor (ER) pathway. ER is a target of current hormonal therapies in breast cancer and improved understanding of the activation of such pathways may be of great use in the development of alternative gene expression-based tumor characterization and treatment.

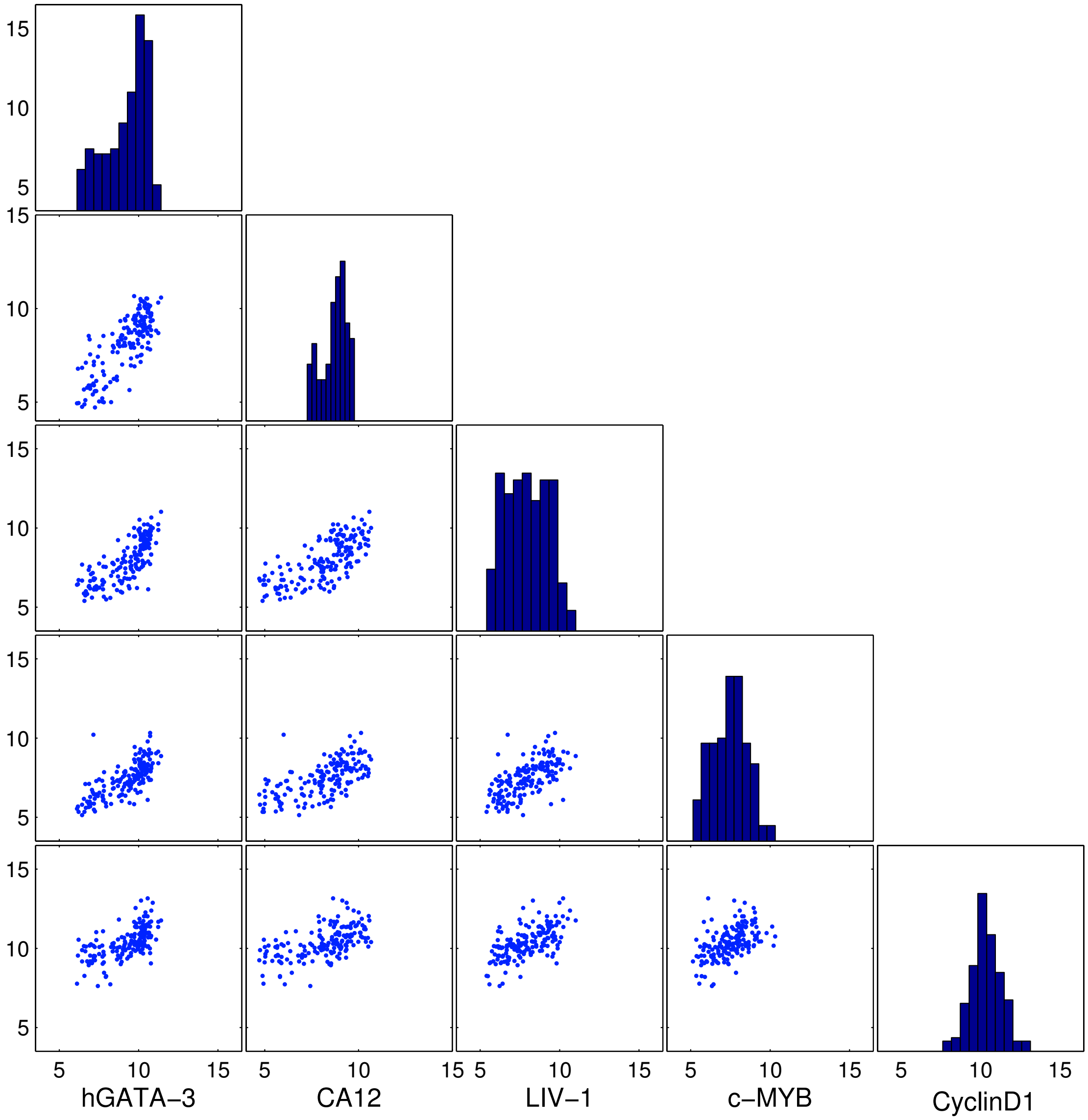


Figure 6.7: Histograms and scatter plots of 5 genes in the initial set of variables for the ER breast cancer example. This data represents log_2 expression estimates processed by RMA.

The evolutionary search started with 10 genes, some directly related to ER (such as hGATA3 and CA12) and others involved in cell cycle activity and growth (Table 6.5.3). Thresholds for inclusion of variables and factors were set at 0.95, with 5 as the minimum number of variables required for the inclusion of a factor. Standard diffuse priors were used for all parameters in the model. To facilitate interpretation, the search was set to stop after 200 variables were included. Figure 6.8 displays the estimated sparsity patterns of **B** providing a visual impression of gene-factor associations across factors as well as cross-talk between factors in terms of genes loaded in more than one factor.

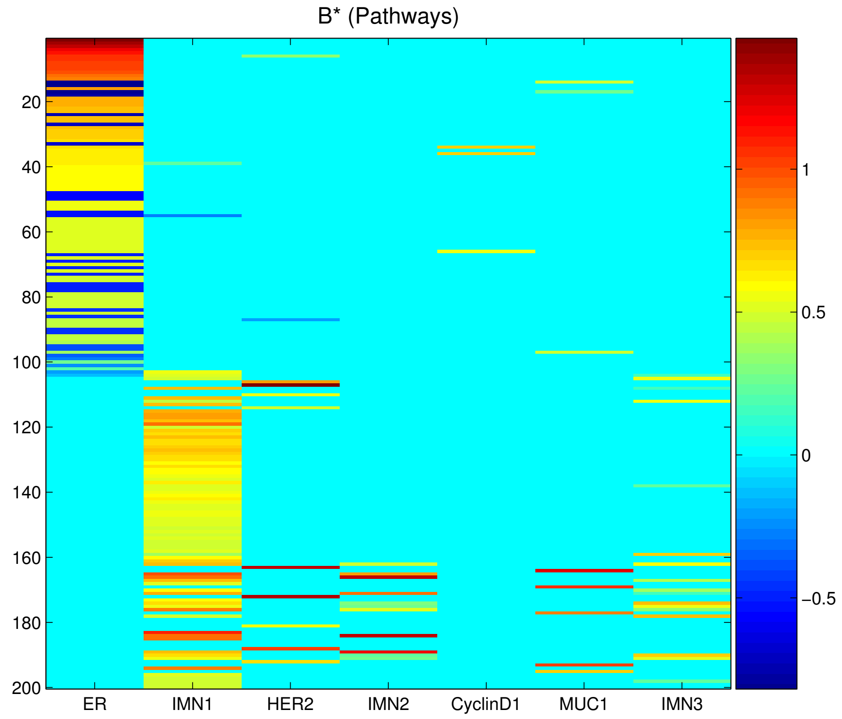


Figure 6.8: Breast cancer gene expression example: Estimate for the loadings matrix in each of the pathways(factors). For $g = 1, \ldots, 200$ and $j = 1, \ldots, 7$, $\beta_{g,j}$ was set to zero if $\pi_{g,j} < 0.95$. The factors are labeled by biological characteristics of its top genes.

By listing the genes loaded on each factor (6.5.3), taking into account the absolute value of the estimated $\beta_{g,j}$, it is possible to examine each subset of genes

hGATA3	Co-expressed with ER (West *et al.*, 2001).
CA12	Over-expressed in malignant breast epithelium (Wykoff *et al.*, 2001).
LIV-1	Estrogen induced (West *et al.*, 2001).
HNF3-α	Synergistic with ER (West *et al.*, 2001).
GRB7	Synergistic with ER (Kristensen *et al.*, 2005).
c-MYB	Estrogen induced (West *et al.*, 2001).
c-MYC	Over-expressed during breast cancer progression (Zelinski *et al.*, 2002).
CyclinD1	Cell cycle regulation and growth(Shoker *et al.*, 2001).
HER2	Over-expressed in aggressive tumors (Yamashita *et al.*, 2004).
ERBB2	Over-expressed in aggressive tumors (Yamashita *et al.*, 2004).

Table 6.5: Starting set of genes and respective functions in the ER breast cancer example.

for common biological functions, allowing us to name the factors and start exploring the biology driving the activity of each pathway. The analysis ends up exploring two large pathways (evident in Figure 6.9), one replete with known ER related genes – the ER pathway – and a second full of genes with immunoregulatory functions related to tumor suppression activity. Genes such as CA12, hGATA3, LIV-1, HNF3-α and c-MYB are highly connected to ER (West *et al.*, 2001) and other oncogenic pathways and play a major role in the estimation of the largest and most dominant factor in the analysis – hence the ER factor. The main immuno-response factor is loaded with genes in the IGL region (Lefranc, 2001), responsible for the production of immunoglobulin which recognizes foreign antigens and initiate immune responses such as phagocytosis. Also in this factor are genes such as Ab63 that has been shown to regulate the proliferation of steam cells and function as a tumor-repressor agent (Cabioglu *et al.*, 2005). RANTES and EBI-1 are other examples of genes involved in tumor-repression loaded in what we call the immuno-response pathway (Moran *et al.*, 2002). The remaining factors clearly reflect other pathways of breast cancer biology (HER2, MUCIN1),

Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
CA12	IGL	ERBB2	V4-31	PRAD1	MUCIN 1	RANTES
hGATA3	Ab63	HER2	IGL	BCL1	PEM	RANTES
LIV-1	Ab63	HER2	V4-31	CyclinD1	PEM	ISGF-3
HNF3-α	IGL	HNF3-α	Ab63		MUCIN	MIR-7
HAG-2/R	RANTES		MB-1		MUCIN	MB-1
AC133	EBI-1		MB-1		C4B	TCF-1
c-MYB	RANTES		PIM-2			CD37
P450	CD45(IGL)					
PDZK1	CD27(IGL)					
EST1866	IGL					

Table 6.6: Breast cancer gene expression example: Symbols for the top genes (probes) loaded in each factor.

cell cycle activity (CyclinD1), and secondary branches of immuno-response activity (IMN2 and IMN3) (Yamashita *et al.*, 2004; Brockhausen *et al.*, 1995; Shoker *et al.*, 2001).

To validate that the factor 1 is in fact an ER factor, I try to predict ER status measured via IHC (immunohistochemistry) proteins assays, reported as ER positive or negative, in each of the tumors using factor 1 as a predictor. IHC assays provide a very imprecise and noisy assessment of ER status but it is, however, the standard way to classify breast tumors. In fact, part of the interest in the pathway exploration analysis is defining a ER factor capable to help improve potential ER assays. Figure 6.10 shows the fitted response probabilities for the logistic regression of ER status on the posterior mean of factor 1 ($\hat{\mathbf{f}}_1$), color coded by the actual ER status. The association of factor 1 and ER is clear.

A final point relates to interpretation of sparse factor models. Figure 6.11 plots the expression levels for the probe CyclinD1 across samples. On the same vertical scale are plotted the fitted values of the two factors for which $\pi_{g,j} > 0.95$

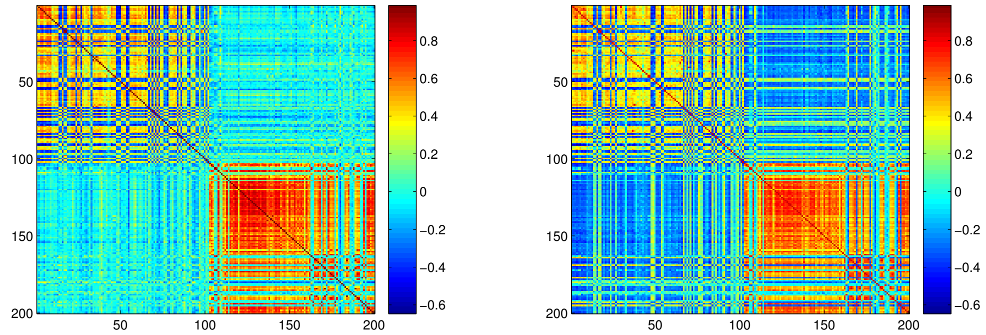


Figure 6.9: ER breast cancer example: Estimate of the correlation matrix (left) and data correlation matrix (right). The two diagonal blocks highlight the activity in the two main pathways explored.

for CyclinD1 and the estimated residuals. The plot illustrates that expression of CyclinD1 can be represented by the activity in two factors across tumors. This provides opportunity for biological interpretation of the participation or the role played by each pathway in the expression of a gene. In this particular example, CyclinD1 is a critical gene in cell cycle regulation, acting to phosphorylate the Rb protein helping cell development and proliferation. Thus some of its variation reflects cell cycle activity, unrelated to ER, represented here by the CyclinD1 factor. However, CyclinD1 is also a direct regulator of c-MYB hence its involvement with the ER pathway. This sort of decomposition generates clear biological rationale for the activity being captured by the factors in this example of a gene of well-known function. It then suggests the ability of such analyses to provide insights and hypotheses about function of other, less well-known genes and aiding in the exploration of important biological pathways.

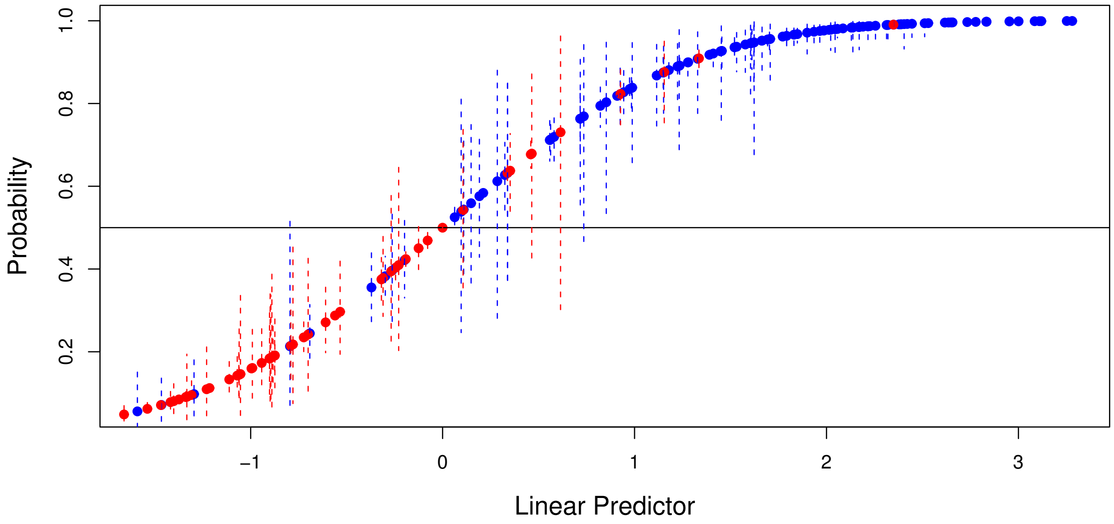


Figure 6.10: Fitted response probabilities for the logistic regression of ER status on the estimated ER factor. The blue points indicate ER positive tumors and the red points indicate ER negative tumors.

6.6 Computational Shortcut

The evolutionary search provides a feasible computational strategy to fit factor models for very many variables. However, when p is very large the variable inclusion step can be very costly as an auxiliary MCMC has to be implemented to rank the many variables outside of the model. In trying to reduce the computational cost of this step we implement an analytical approximation that replaces the auxiliary MCMC and provide a much faster way to perform the inclusion step. Given that we are interested in bringing into the model variables highly associated with the current factors, the simplest approximation would be to rank each variable x_g outside of the model by the R^2 of the regression on $\hat{\mathbf{F}}$. This, however, doesn't take

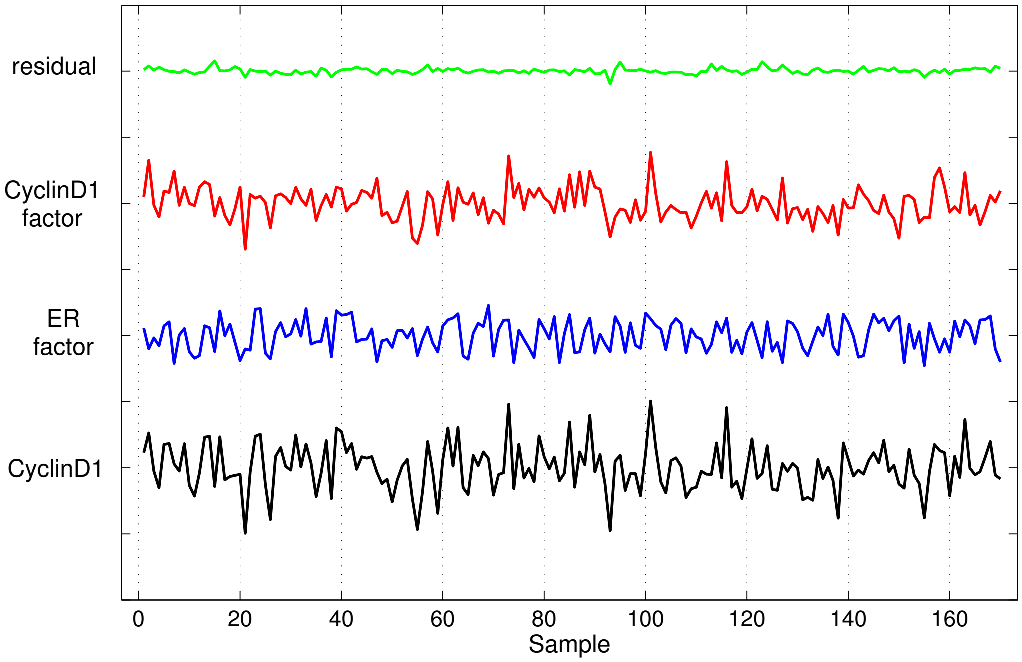


Figure 6.11: Plot across breast tumors samples of CyclinD1, the ER factor and the CyclinD1 factor. All plots are in the same vertical scale, indicating the breakdown of the expression variation of CyclinD1 in two components.

into account the sparsity inducing priors and creates a decision rule that is not based on the inclusion probability $\pi_{g,j}$. Our approximation instead estimates the inclusion probability $\pi_{g,j}$ for each factor j ($j = 1, \ldots, k$) in a univariate regression. The variables are then ranked by their highest inclusion probability on any of the current k factors.

Let x_g be one of the variables out of the model and $\hat{\mathbf{F}}_j$ be the estimated scores for factor j. The posterior probability of $\beta_{g,j} \neq 0$ can be approximated as

$$\hat{\pi}_{g,j} \approx \frac{\hat{\rho}_j m_j}{\hat{\rho}_j m_j + (1 - \hat{\rho}_j m_j) N(0|\hat{\mu}_{g,j}, \hat{C}_{g,j}) / N(0|0, \hat{\tau}_j)} \tag{6.26}$$

where

$$\hat{C}_{g,j} = \left(\frac{1}{\hat{\tau}_j} + \frac{\hat{\mathbf{F}}'_j\hat{\mathbf{F}}_j}{\hat{\psi}_g}\right)^{-1}, \tag{6.27}$$

$$\hat{\mu}_{g,j} = \hat{\psi}_g^{-1}\hat{C}_{g,j}\hat{\mathbf{F}}'_j x_g, \tag{6.28}$$

and

$$\hat{\psi}_g = \frac{\sum_{i=1}^{n}(x_{g,i} - \hat{\beta}_{g,j}\hat{f}_{j,i})^2 + b_\psi}{n + a_\psi - 1} \tag{6.29}$$

with $\hat{\beta}_{g,j}$ being the least squares estimate of $\beta_{g,j}$. For a given $\hat{\tau}_j$ (from the previous MCMC) and a $\hat{\rho}_j$ (discussed below) this approximation is obtained by a normal approximation to $p(x_g|\beta_{g,j} \neq 0)$, the marginal likelihood of x_g, after integrating out ψ_g.

Note that using ρ_j from the previous MCMC can lead to overestimation of the inclusion probability of many variables as at this point ρ_j has been estimated with the number of variables currently in the model. If most of the variables in the model are loaded in factor j, estimates of ρ_j will take large values which, if used in the approximation, will then artificially increase the propensity of variables to load on factor j. We correct this by adjusting the current estimate of ρ_j by the prior proportion of inclusion for all p_{out} variables out of the model, leading to

$$\hat{\rho}_j = \frac{rs + p_{load} + p_{out}rs}{s + p_{in} - j + p_{out}} \tag{6.30}$$

where p_{in} is the number of variables currently in the model, p_{load} are the number of variables estimated to be loaded on factor j with r and s being the hyperparameters of the prior for ρ_j.

Experimenting with this approximation in the simulated dataset presented in Section 6.5.1 generated very similar results with the running time being significantly reduced. In all but 12 runs (out of 100) every relevant variable was included and in every run 7 was the final number of factors. Again, the variables left out were part of the same group displayed in Table 6.5.1.

6.7 Discussion

The main contribution of this chapter is the development of an evolutionary search concept and the implied method of sequentially exploring specific components of variation in high-dimensional datasets. Sparsity of the loadings matrix $\mathbf{B}$ is the fundamental idea that guides the search by providing a formal way to identify variables related to the latent structure and help check the need for new factors. As described here, this work was motivated by the exploration of oncogenetic pathways but its applicability is broad as highlighted in the financial example presented. This procedure has also proved to be very useful in variable selection for latent factor regression as developed in Carvalho *et al.* (2005) where multiple response variables are jointly modeled with a high-dimensional set of predictive variables.

A general software tool implementing the models and methods presented here is currently being developed and will be available on-line very soon. BFRM (Bayesian Factor Regression Model) is configured to run MCMC analysis of sparse factor models and factor regression models including the evolutionary search. BFRM is available in a 32 and 64-bit version compatible with both Linux and Windows operational systems. A multi-threaded version of the software is currently under development. This project is a collaboration with Quanli Wang and Mike West, part of the Duke Integrated Cancer Biology Program (ICBP). Details of BFRM are presented in Appendix B.

Chapter 7

Final Comments and Extensions

7.1 Summary

This dissertation addressed a range of important issues in sparse models for large-scale multivariate problems. The discussion focused on models for covariance and precision matrices with Gaussian graphical models and sparse factor models as parsimonious structures for representing complex, high-dimensional relationships in terms of simpler, lower-dimensional structures.

The research reported develops theoretical and methodological aspects of high-dimensional multivariate problems, and innovative computational tools for model selection and inference. In Gaussian graphical models I presented a novel *shotgun stochastic search* for model selection, developed an efficient sampling scheme for the HIW, and extended conditional independence ideas to time series analysis by defining a new class of multivariate dynamic linear models. In sparse factor models, a key contribution is the development of an *evolutionary search* that addresses important questions of model specification, variable identification and hard practical issues of mapping substructure in very high-dimensional problems.

Some of the work in this dissertation was motivated by large-scale genomics studies. As demonstrated in the financial examples, however, the concept of sparsity is vast in its applicability and of potentially key relevance in its practical import.

As scientific problems continue to grow in dimension, extensions of the ideas and methods discussed here are a pressing need. With increasing access to larger clusters for distributed computing, statistics research has an opportunity to substantially advance our ability to explore complex, high-dimensional model spaces by integrating technological advances into theoretical and methodological research goals.

I now conclude this dissertation by listing some extensions to the work presented.

7.2 Extensions in Gaussian Graphical Models

From the results in Chapter 3 it is obvious that much has to be done in order for non-decomposable graphs to be routinely considered in high-dimensional situations. The development of theoretical insights and methods are necessary to improve the capacity to estimate the normalizing constants associated with non-complete prime components. One potential direction for research is to understand the changes in the junction tree of non-decomposable graphs when one-edge moves are considered. Flores *et al.* (2003) addressed this problem in the context of direct graphs and adapting their results to prime component changes in undirected graphs could lead to simplifications in line with what is described in Section 3.3 of chapter 3 and Appendix A. In the same direction, creating a map from non-decomposable to the space of decomposable graphs might generate a good, approximate way to explore the entire space of graphs by working around decomposable graphs.

Recently, a rather different view of graphical model search has been outlined in Dobra *et al.* (2004). These are *constructive* methods where the full joint distribution is derived using a triangular set of regressions representing the relationships between variables. This is related to both the dependency network framework of Heckerman *et al.* (2000) and approaches that model structure in the Cholesky decomposition of variance matrices; it is innovative in the creation of an approach that scales with dimension, encourages graph sparsity, and utilizes priors consistent across graphs. These methods can handle large sets of variables due to the pre-screening procedure that limits which variables are considered possible predictors of others. This type of constructive method generates graphs that are potentially widely different at each step, especially if compared to the one-edge move strategy described in Chapter 2. Understanding the connections and theoretical differences between this approach and the undirected Gaussian graphical models is necessary as these methods are able to analyze problems of thousands of variables.

In this dissertation I have only considered model selection in Gaussian graphical models. Graphs, however, are very useful tools for other models – contingency tables for example (Lauritzen, 1996) – and extensions of SSS to other contexts is another important direction for future research.

7.3 Portfolio Problems

In Chapter 5, a very general multivariate dynamic linear model with structured covariance matrix was developed. In applying this model to large-scale portfolio problems we found that conditional independence constraints help reduce investment uncertainty and potentially generate more profitable opportunities. An intuitive interpretation of this result is given through Equation 5.17 where we see that the variance of optimal portfolio weights increases if zero constraints in the precision matrix are ignored. High-dimensional portfolios are regarded as one of the most challenging problems in financial theory (Polson and Tew, 2000) and theoretical developments for a precise understanding of the connections between conditional independence assumptions and optimal investment strategies is of key importance for further advances in that area. Still in this context, I hope to explore the performance of DLMs with graphical constraints in more realistic applications of portfolio problems (e.g. Quintana, 1992; Quintana *et al.*, 1995) where transaction costs and other economic variables are considered.

The fact that sparsity modeling of the precision matrix of assets has shown to be very relevant in portfolio problems should not be unique to dynamic linear models. Therefore, another important research direction in this area is to explore the impact of graphical constraints in different types of dynamic covariance models – such as multivariate stochastic volatility models (Aguilar, 1998; Chib *et al.*, 2004) and dynamic conditional correlations models (Engle, 2002).

A final question in this area is the development of efficient sequential model selection procedures for graph identification. Multi-process models (class I), as described in Section 5.6, provide a nice and easily parallelizable strategy to account for model uncertainty but its performance is limited to how representative is the set of graphs in the mixture within the entire space of models. Treating the graph as a state and including its update in the sequential estimation might be one possible alternative for this problem. This could be implemented via a particle filter strategy where at each step the set of "particles" (graphs) are updated according to some evolution. The nice feature of this proposal is the fact that, for a given set of particles, computations could be efficiently implemented in parallel.

7.4 Future Work in Sparse Factor Models

As a continuation of the effort to identify and understand pathways of biological activity, we will continue to explore genome-wide expression datasets with sparse factor models. Part of the work currently being developed at the Duke ICBP focuses on predictive models for phenotypes based on pathways identified by factors (Carvalho *et al.*, 2005). The evolutionary search, in this case, works as a variable selection procedure that refines the exploration of the predictive pathways while identifying genes that have a direct impact on the responses. The development of software for the implementation of Bayesian factor regression models is also part of ICBP's effort and a brief description of BFRM appears in Appendix B.

As I have shown, the usefulness of exploratory methods such as the evolutionary search is clear. Constant testing of its performance in simulated and real problems has been very satisfactory. However, it is also clear that further theoretical developments to expand our understanding of search methods in factor models space are needed. Trying to embed the evolutionary search in a formal model selection paradigm – exhibiting its role and relationship to MCMC "search" over a global model involving all variables – is part of my near-term research agenda.

7.4.1 Matric-Variate Factor Models

One possibly interesting research area is the extension of factor models to matrices. As an example, consider multiple economic indicators being observed in a collection of different countries across time. This is the case in many longitudinal studies and one might be interested in understanding the underlying structure among countries, indicators and across them. This type of data could be modeled through what we call Matrix Factor Models where for the i^{th} observation

$$\mathbf{Y}_i = \mathbf{A}\mathbf{\Lambda}_i\mathbf{B} + \mathbf{E}_i \tag{7.1}$$

with the following assumptions:

- $\mathbf{Y}$ is a $(p \times q)$ random matrix;
- $\mathbf{\Lambda}_i$ in a $(k \times h)$ matrix of factor scores with prior $\mathbf{\Lambda}_i \sim N(\mathbf{0}, \mathbf{I}_k, \mathbf{I}_h)$;
- $\mathbf{A}_{p \times k}$ and $\mathbf{B}_{h \times q}$ are factor loadings matrices;
- $\mathbf{E}_i$ is the matrix of idiosyncratic noise following a $N(\mathbf{0}, \mathbf{I}_p, \mathbf{I}_q)$.

This model implies that

$$vec(\mathbf{Y}) \sim N(0, (\mathbf{A}\mathbf{A}' + \mathbf{I}_p) \otimes (\mathbf{B}\mathbf{B}' + \mathbf{I}_q))$$

which is equivalent to the composition of two separate factor models, one for the rows and another for the columns of $\mathbf{Y}$. Fitting the model in (7.1) should be just an extension of the MCMC described in Chapter 6. Questions of model specification are again the complicated part and central for the development of methods necessary for the implementation of such models.

7.4.2 Sparse Factor-Graphical Models: A synthesis?

One further, very interesting set of questions – an an open research direction that may yield a very promising agenda – concerns the relationships between sparse factor models and (also sparse) graphical models. At a more general level this question has to do with the complementarities of approaches to modeling sparse structures through the conditional independence approach (graphical models) and the association or dependence approach (factor models). Little seems to have been done to understand and reconcile what might appear to be conflicting approaches but should in fact be complementary and consistent.

In the Gaussian case as discussed in this thesis, one focusing question concerns the relationships between the structure and sparsity of a factor loadings matrix $\mathbf{B}$ and the implied structure of the graph, equivalently the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$. Sparse factor models induce many zeros in $\mathbf{B}$, and so impose constraints on $\mathbf{\Sigma}$ directly; inverting to the precision matrix will generally reduce the sparsity – that is, a very sparse $\mathbf{\Sigma}$ (factor model, association or dependence graph implied) will generally invert to $\mathbf{\Omega}$ that – at least in terms of just the simple proportion of zero elements – is much less sparse. Inversely, beginning with a very sparse graph, hence a very sparse $\mathbf{\Omega}$, inversion can lead to a much less sparse $\mathbf{\Sigma}$. This raises questions about moving between the two approaches and of reconciling sparsity modeling in the two views. Understanding the connections between the two approaches might generate new ideas and expand our ability to model high-dimensional covariance structure. One aspect of this general set of questions and an attempt to establish this connection appears in Jones and West (2005); there new theory is defined to represent the covariance between two variables in terms of a decomposition over a graph - a covariance is decomposed into a sum of "path weights" for all paths connecting the two variables in an undirected graph, and these path weights relate intimately to the elements of the precision matrix. Conditions on the precision matrix that "zero out" any given covariance would then aid in reconciling the two representations.

In trying to better understand the connections between $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ in a sparse factor model context, we note the following intriguing identity and possible lead in a new research direction. Assume that $\mathbf{x}$ is modeled via a factor model as defined in Equation (6.1) with $\mathbf{f}_i \sim N(\mathbf{0}, \mathbf{Q})$ and $\boldsymbol{\nu}_i \sim N(\mathbf{0}, \mathbf{I})$. This implies that

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{Q}\mathbf{B}' + \mathbf{I}. \tag{7.2}$$

Now, we can show that if $\mathbf{Q}$ is given by

$$\mathbf{Q} = g^{-1}(\mathbf{B}'\mathbf{B})^{-1}$$

for some constant $g > 0$, then the precision matrix $\mathbf{\Omega}$ is given by

$$\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = \mathbf{\Omega} = \mathbf{I} - (1+g)^{-1}\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'. \tag{7.3}$$

This result utilizes the fact that $\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$ is an idempotent matrix.

One consequence is that the sparsity pattern – in terms of the positions of off-diagonal zeros – in the covariance matrix and the precision is precisely the same. Hence we can generate models in which there is a fundamental sparsity pattern – hence one graph – shared by the correlation/association and the precision matrix simultaneously. From the above, we can show that the specific "idempotent" graph structure is a direct consequence of the sparse factor model in which the prior for the latent factors is modified to a prior precision matrix $\mathbf{Q}^{-1} = g(\mathbf{B}'\mathbf{B})$. This is essentially the use of a standard Bayesian $g-$prior (Zellner, 1986; West, 2003) for $\mathbf{f}_i$, generating a posterior for the latent factors that have the same covariance structure as observed (given $\mathbf{B}$) in the likelihood. (The above is developed relative to a model in which the idiosyncratic variances matrix $\mathbf{\Psi} = \mathbf{I}$; in the more general case, the extension is simply $\mathbf{Q}^{-1} = g(\mathbf{B}'\mathbf{\Psi}^{-1}\mathbf{B})$ and the result about the common sparsity pattern in covariance and precision graphs is maintained).

So, the knowledge of the covariance structure of the latent factors allow the creation of a new class of models that establishes a link between factors and Gaussian graphical models: if we view graphs as fundamental, this class of models is a general class in which graphical models (defined by zeros in precision) and association graphs (zeros in covariance matrices) coincide and so are reconciled. The further pursuit of these ideas is yet another subject of my expected near-term research.

Appendix A

Graphs: Results and Algorithms

A.1 Maximum Cardinality Search and Decomposable Graphs

In this section we will consider how to obtain a junction tree representation of a connected decomposable graph. To obtain a junction forest of a disconnected graph, the algorithm can be used on each connected component. Obtaining this representation for non-decomposable graphs builds on this algorithm and is considered in Section A.1.1. The junction forest is created by first establishing a perfect ordering of the nodes of the graph, using the following maximum cardinality search algorithm:

1. Pick a vertex v, and label it 1. While some unlabeled vertices remain, iterate the following procedure:

2. Suppose k unlabeled vertices remain. From among the vertices with the most labeled neighbors, pick a vertex and label it $p - k + 1$.

One can use this algorithm to check for decomposability of a graph by checking at each iteration of step (2) that all the labeled neighbors of the vertex to be added form a complete subgraph.

For decomposable graphs, the ordering of vertices established defines an ordering of cliques, where the cliques are ordered by the highest numbered node contained in each. This sequence has the running intersection property: for all $j > 1$, let S_j be the set of nodes shared with lower numbered cliques. There is an $i < j$ such that $S_j \subset C_i$, and the S_j's are all complete. Thus S_j is a separator between $C_1, \ldots, C_{j-1}$ and $C_j \ldots C_k$. This property shows us that the cliques can be arranged in a junction tree, where cliques are nodes, and cliques that share vertices are connected by an edge. Clique C_j may contain the separators of and therefore be connected to many higher numbered cliques, but it is connected to at most one lowered number clique. This prevents loops in the connections among cliques, telling us the structure is a tree. The highest numbered clique is a leaf, connected to only one other clique. While there may be many perfect orderings (for examples, leaves of the tree may be listed in any order among themselves) the junction tree is a unique representation.

A.1.1 Non-decomposable Graphs

Non decomposable graphs also have a junction forest representation, but in terms of the prime components $P_1 \ldots P_k$ rather than cliques. To get at this representation, we first triangulate the graph (add edges so that it is decomposable). A perfect ordering is then built as in Section A.1. The set of edges added during triangulation are called the *fill-in* edges. Now we will remove the fill-in edges and consolidate the prime components that were decomposed after the addition of these edges, while maintaining the running intersection property in our ordering of prime components. Any of the fill in edges not in $S_2, \ldots, S_k$ can simply be removed. To deal with the other edges, we start with the highest numbered separator S_j containing fill-in edges. We consolidate C_j and the lower numbered component containing S_j, C_i. The sequence of cliques then reads $C_1, \ldots C_{j-1}, C_{j+1}, \ldots C_k$. This maintains the running intersection property–any separators contained in C_j are now contained in the lower numbered clique C_i. We repeat this process in sequence for each separator containing fill-in edges.

A.2 One edge changes that maintain decomposability

It has long been known that an edge deletion maintains decomposability if that edge is contained in exactly one clique (see, for example, Frydenberg and Lauritzen 1989). Giudici and Green (1999) give an efficient condition for checking whether an edge addition maintains decomposability. Decomposability is maintained if the vertices to be joined (a and b) are in different connected components or if there exist $R, T \subset V$ such that $a \cup R$ and $b \cup T$ are cliques, and $S = R \cap T$ is a separator on the path between $a \cup R$ and $b \cup T$ in the junction forest representation of the graph G. In our program, the junction forest representation of the graph is maintained, listing the cliques and separators of each component. When considering adding an edge between a and b in the same component, each possible combination of values of R and T are considered (these are defined by the clique memberships of a and b). For each of these combinations, it is determined whether $R \cap T$ is a separator. As demonstrated in Giudici and Green (1999), checking these conditions results in substantial time savings over checking the decomposability of the new graph with maximum cardinality search each time. Other conditions for checking whether edge addition maintains decomposability are given in Deshpande *et al.* (2001); however we found them more difficult to implement in practice.

Appendix B

BFRM: Bayesian Factor Regression Model

BFRM is software developed for the implementation of sparse Bayesian factor regression models in line with what was presented in Chapter 6. BFRM will soon be made available at *www.isds.duke.edu* under the software link.

BFRM fits the following general model (described in details in Carvalho *et al.*, 2005):

$$\begin{pmatrix} \mathbf{y}_i \\ \mathbf{x}_i \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{B} & \mathbf{C} \\ \mathbf{D} & \mathbf{E} & \mathbf{F} \end{pmatrix} \begin{pmatrix} \mathbf{H}_i \\ \boldsymbol{\mu}_i \\ \boldsymbol{\lambda}_i \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_i \\ \boldsymbol{\nu}_i \end{pmatrix} \tag{B.1}$$

where

- $\mathbf{y}_i$ is the response q-vector (continuous, binary, categorical or possibly censored survival data); $\mathbf{x}_i$ is the p-vector of candidate predictive variables (continuous variables);
- $\boldsymbol{\mu}_i$ is the latent response factor q-vector; $\boldsymbol{\lambda}_i$ is the latent factor k-vector; $\mathbf{H}_i$ is a h-vector of regressors (and control) variables;

- $\mathbf{A}_{q\times h}$ and $\mathbf{D}_{p\times h}$ are regression coefficient matrices relating both $\mathbf{y}_i$ and $\mathbf{x}_i$ to $\mathbf{H}_i$;
- $\mathbf{B}_{q\times q}$ is the matrix of response factor loadings linking the response variables to response factors;
- $\mathbf{E}_{p\times q}$ is the matrix of factor loadings relating the response factors to $\mathbf{x}_i$ whereas $\mathbf{F}_{p\times k}$ is the factor loadings matrix;
- $\mathbf{C}_{q\times k}$ is the factor loadings matrix relating the responses to the latent factors $\boldsymbol{\lambda}_i$;
- $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\nu}_i$ are the idiosyncratic noise term.

All factor loading matrices and matrix of regression coefficients are expected to be sparse thus modeled through the sparsity priors described in Chapter 6.

This model represents the view that relationships between the response vector and $\mathbf{x}$ can be direct or implicit through the latent factor structure.

BFRM is also able to run an evolutionary search to expand the sample space of $\mathbf{x}$ sequentially including variables to enrich the estimation of latent factors $\boldsymbol{\lambda}$ and improve the predictive ability of the model.

Another important feature of BFRM is its ability to handle missing observations in the response vector $\mathbf{y}$ which are imputed in the MCMC.

An example of the parameter file used as input for BFRM follows:

```
#Version 1.5 (January 18th 2006)

#data section
NObservations = 1000           # number of observations
NVariables = 200               # number of X variables
```

```
NBinaryResponses = 0           # number of binary responses
NCategoricalResponses = 0      # number of categorical responses
NSurvivalResponses = 0         # number of survival responses
NContinuousResponses = 0       # number of continuous responses
NDesignVariables = 1           # number of regressor (size of H)
NLatentFactors = 2             # number of latant factor k (starting point)
DataFile = dataset.txt         # X Data file (All X’s)
HFile    = H.txt               # H Data file (Regressors)
ResponseMaskFile = YMask.txt   # Indicator of missing observations in Y

#prior section
#prior Psi
PriorPsia = 10
PriorPsib = 2
PriorSurvivalPsia = 2
PriorSurvivalPsib = 0.5
#prior Rho
PriorRhoMean = 0.001
PriorRhoN = 200
#prior Pi
PriorPiMean = 0.9
PriorPiN = 10
#prior Tau (Possibly different for each response factor)
PriorTauDesigna = 5
PriorTauDesignb = 1
PriorTauResponseBinarya = 5
PriorTauResponseBinaryb = 1
PriorTauResponseCategoricala = 5
PriorTauResponseCategoricalb = 1
PriorTauResponseSurvivala = 5
PriorTauResponseSurvivalb = 1
PriorTauResponseContinuousa = 5
```

```
PriorTauResponseContinuousb = 1
PriorTauLatenta = 5
PriorTauLatentb = 1

#evolutionary mode section
Evol = 1                   # Option for evolutionary search 0/1 (no/yes)
EvolVarIn = 20             # number of variables in the initial model
EvolVarInFile = varin.txt  # file indicating what variables are in
# Probability threshold for variable inclusion
EvolIncludeVariableThreshold = 0.95
# Probability threshold for factor inclusion
EvolIncludeFactorThreshold = 0.8
# minimun number of variables needed for factor inclusion
EvolMiniumVariablesInFactor = 3
EvolMaximumFactors = 50               # Maximum number of factors
EvolMaximumVariables = 100            # Maximum number of variables
# Maximum number of variables included per iteration
EvolMaximumVariablesPerIteration = 10

#mcmc section
Burnin = 2000
Burnin_Select = 1000     # Burnin for auxiliary  MCMC
nMCSamples = 2000
nMCSamples_Select = 2000 # Monte Carlo samples for auxiliary MCMC

#monitoring section
PrintIteration = 100

DEBUG = 0
CheckPoint = 0
# Use of approximation in variable inclusion step 0/1 (no/yes)
InclusionMethod = 1
```

Appendix C

Multivariate and Matrix Distributions

C.1 The Matrix Normal Distribution

A random matrix $\mathbf{X}_{q\times p}$ is said to follow a Matrix Normal Distribution (Dawid, 1981) denoted by $N(\mathbf{M}, \mathbf{C}, \mathbf{\Sigma})$, with mean $\mathbf{M}_{q\times p}$, left covariance matrix $\mathbf{C}_{q\times q}$ and right covariance matrix $\mathbf{\Sigma}_{p\times p}$ if its density is given by

$$p(\mathbf{X}) = (2\pi)^{-qp/2}|\mathbf{C}|^{p/2}|\mathbf{\Sigma}|^{q/2}exp\left\{-\frac{1}{2}tr\left[(\mathbf{X}-\mathbf{M})'\mathbf{C}^{-1}\mathbf{X}\mathbf{\Sigma}^{-1}\right]\right\}. \tag{C.1}$$

Some important properties of the Matrix Normal Distribution are given below:

(i) $vec(\mathbf{X}) \sim N(vec(\mathbf{M}), \mathbf{C}\otimes\mathbf{\Sigma})$ where $vec(\mathbf{X})$ is the usual column-vectorization of the matrix $\mathbf{X}$;

(ii) $\mathbf{X}' \sim N(\mathbf{M}', \mathbf{\Sigma}, \mathbf{C})$;

(iii) For any matrix $\mathbf{H}_{q\times q}$, $\mathbf{K}_{p\times p}$ and $\mathbf{L}_{q\times p}$

$$\mathbf{HXK}+\mathbf{L} \sim N(\mathbf{HMK}+\mathbf{L}, \mathbf{HCH}', \mathbf{K}'\mathbf{\Sigma}\mathbf{K}); \tag{C.2}$$

(iv) Marginal and conditional distribution for any elements of $\mathbf{X}$ are normal distributed. Without loss of genererality, consider marginalization and conditioning by rows, where

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

so that

$$\mathbf{X}_1 \sim N(\mathbf{M}_1, \mathbf{C}_{11}, \mathbf{\Sigma}) \qquad \text{and} \qquad \mathbf{X}_{2|1} \sim N(\mathbf{M}_{2|1}, \mathbf{C}_{2|1}, \mathbf{\Sigma}) \tag{C.3}$$

with

$$\begin{aligned} \mathbf{M}_{2|1} &= \mathbf{M}_2 + \mathbf{C}_{21}\mathbf{C}_{11}^{-1}(\mathbf{X}_1 - \mathbf{M}_1) \\ \mathbf{C}_{2|1} &= \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12} \end{aligned}$$

C.2 The Matrix T Distribution

The Matrix T Distribution is a analogue of the multivariate T defined in Dawid (1981). Let the columns of $\mathbf{X}_{q\times p}$ follow a multivariate T distribution with n degrees of freedom and write $\mathbf{X} = (X_1, \ldots, X_p)$ and $\mathbf{m} = (m_1, \ldots, m_p)$ so that

$$X_i \sim T_n(m_i, \mathbf{C}S_i) \qquad i = 1:p$$

with m_i and $\mathbf{C}S_i$ as the location and scale parameter respectvely. As with the matrix normal distribution notation, $\mathbf{X}$ follows a matrix T distribution denoted by:

$$\mathbf{X} \sim T_n(\mathbf{m}, \mathbf{C}, \mathbf{S}), \tag{C.4}$$

with density given by

$$p(\mathbf{X}) = k|\mathbf{C}|^{-p/2}|\mathbf{S}|^{-q/2}|\mathbf{I}_q + n^{-1}[\mathbf{C}^{-1}(\mathbf{X}-\mathbf{m})][(\mathbf{X}-\mathbf{m})\mathbf{S}^{-1}]'|^{-(n+q+p-1)/2} \tag{C.5}$$

with normalizing constant k defined as

$$k = (n\pi^2)^{-(pq)/2} \frac{\Gamma_{p+q}\left(\frac{n+p+q-1}{2}\right)}{\Gamma_p\left(\frac{n+p-1}{2}\right)\Gamma_q\left(\frac{n+q-1}{2}\right)}$$

with Γ denoting the multivariate gamma function.

Bibliography

Aguilar, O. (1998). *Latent Structure in Bayesian Multivariate Time Series Models*. PhD. Thesis, Duke University.

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and variance matrix discounting for portfolio allocation. *Journal of Business and Economic Statistics* **18**, 338–357.

Ameen, J. and Harrison, P. (1985). Normal discount Bayesian models. In J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, eds., *Bayesian Statistics II*, 271–298. Valencia University Press.

Arminger, G. and Múthen, B. (1998). A Bayesian approach to nonlinear latent variable models using Gibbs sampler and the Metropolis-Hastings algorithm. *Psycometrika* **63**, 271–300.

Atay-Kayis, A. and Massam, H. (2005). The marginal likelihood for decomposable and non-decomposable graphical Gaussian models. *Biometrika* **92**, 317–335.

Bartholomew, D. (1984). The foundations of factor analysis. *Biometrika* **71**, 221–232.

Bartlett, M. S. (1933). On the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh* **53**, 260–283.

Bollerslev, T., Chou, R., and Kroner, K. (1992). ARCH modeling in finance. *Journal of Econometrics* **52**, 5–59.

Brockhausen, I., Yang, J., Burchell, J., Whitehouse, C., and Taylor-Papadimitriou, J. (1995). Mechanisms underlying aberrant glycosylation of muc1 mucin in breast cancer cells. *European Journal of Biochemistry* **233**, 607–617.

Cabioglu, N., Yazici, M., Arun, B., Broglio, K., Hortobagyi, G., Price, J., and Sahin, A. (2005). CCR7 and CXCR4 as novel biomarkers predicting axillary lymph node metastasis in t1 breast cancer. *Clinical Cancer Research* **11**, 5686–5693.

Carvalho, A. and Tanner, M. (2005). Mixtures-of-experts of autoregressive time series: Asymptotic normality and model specification. *IEEE Transactions on Neural Networks* **16**, 39–56.

Carvalho, C., Wang, Q., Lucas, J., Chang, J., Nevins, J., and West, M. (2005). High-dimensional sparse factor models and latent factor regression. ISDS Discussion Paper.

Chib, S., Nardari, F., and Shephard, N. (2004). Analysis of high dimensional multivariate stochastic volatility models. Preprint.

Clyde, M., DeSimone, H., and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association* **91**, 1197–1208.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science* **19**, 81–94.

Dawid, A. P. (1980). Conditional independence for statistical operations. *The Annals of Statistics* **8**, 598–617.

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

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics* **3**, 1272–1317.

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

Dellaportas, P., Giudici, P., and Roberts, G. (2003). Bayesian inference for nondecomposable graphical Gaussian models. *Sankhya, Series A* **65**, 43–55.

Dempster, A. (1972). Covariance selection. *Biometrics* **28**, 157–175.

Deshpande, A., Garofalakis, M. N., and Jordan, M. I. (2001). Efficient stepwise selection in decomposable models. In J. Breese and D. Koller, eds., *Uncertainty in Artificial Intelligence (UAI), Proceedings of the Seventeenth Conference.*

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Annals of Mathematical Statistics* **42**, 204–223.

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

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics* **20**, 339–350.

Flores, M. J., Gamez, J., and Olesen, K. G. (2003). Incremental compilation of Bayesian networks. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, 233–240, San Francisco, CA. Morgan Kaufmann Publishers.

Friedman, J. and Tukey, J. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions and Computers* **c23**, 881–889.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistics Association* **85**, 398–409.

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

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies* **9**, 557–587.

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

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

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

Grone, R., Johnson, C. R., Sà, E. M., and Wolkowice, H. (1984). Positive definite completions of partial hermitian matricies. *Linear algebra and its applications* **58**, 109–124.

Hammersley, J. M. and 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. Thesis, Duke University.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective.* Springer, New York.

Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., and Kadie, C. (2000). Dependency networks for inference, collaborative filtering, and data visualization. *Journal Of Machine Learning Research* **1**, 49–75.

Huang, E., Cheng, S., Dressman, H., Pittman, J., Tsou, M.-H., Horng, C.-F., Bild, A., Iversen, E., Liao, M., Chen, C.-M., West, M., Nevins, J., and Huang, A. (2003). Gene expression predictors of breast cancer outcomes. *Lancet* **361**, 1590–1596.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* **31**, e15.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **2**, 249–264.

Jacquier, E., Polson, N., and Rossi, P. (1994). Bayesian analysis of stochastic volatility models. *Journal of Business ans Economic Statistics* **12**, 371–417.

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

Jones, B. and West, M. (2005). Covariance decomposition in undirected graphical models. *Biometrika* **92**, 779–786.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH model. *Review of Economic Studies* **65**, 361–393.

Kristensen, V., Sorlie, T., Geisler, J., Langerod, A., Yoshimura, N., Karesen, R., Harada, N., Lonning, P., and Borresen-Dale, A. (2005). Gene expression profiling of breast cancer in relation to estrogen receptor status and estrogen-metabolizing enzymes: Clinical implications. *Clinical Cancer Research* **11**, 878–883.

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

Lawley, D. and Maxwell, A. (1971). *Factor Analysis as a Statistical Method*. Butterworths, London.

Ledoit, O. and Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management* **30**, 110–119.

Lefranc, M.-P. (2001). Nomenclature of the human immunoglobulin lambda (IGL) genes. *Experimental and Clinical Immunogenetics* **18**, 242–254.

Liu, J. (2000). *Bayesian Time Series Analysis: Methods Using Simulation-Based Computation*. PhD. Thesis, Duke University.

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

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2006). Sparse statistical modelling in gene expression genomics. *Bayesian Bioinformatics* .

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

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

Martin, J. and McDonald, R. (1975). Bayesian estimation in unrestricted factor analysis: a treatment for heywood cases. *Psychometrika* **40**, 505–517.

Massam, H. and Neher, E. (1998). Estimation and testing for lattice conditional independence models on euclidean jordan algebras. *The Annals of Statistics* **26**, 1051–1082.

Moran, C. J., Arenberg, D., Huang, C., Giordano, T., Thomas, G., Misek, D., Chen, G., Iannettoni, M., Orringer, M., Hanash, S., and Beer, D. (2002). RANTES expression is a predictor of survival in stage i lung adenocarcinoma. *Clinical Cancer Research* **8**, 3803–3812.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York: Wiley.

Perold, A. (1988). Large-scale portfolio optimization. *Management Science* **30**, 1143–1160.

Pittman, J., Huang, E., Dressman, H., Horng, C., Cheng, S., Tsou, M., Chen, C., Bild, A., Iversen, E., Huang, A., Nevins, J., and West, M. (2004). Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences* **101**, 8431–8436.

Polson, N. and Tew, B. (2000). Bayesian portfolio selection: An empirical analysis of the S&P500 index 1970 - 1996. *Journal of Business and Economic Statistics* **18**, 164–173.

Press, S. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference (2nd edition)*. New York: Krieger.

Press, S. and Shigemasu, K. (1989). Bayesian inference in factor analysis. In *ASA Proceedings of Social Statistics Section*, 292–294.

Putnam, B. and Quintana, J. (1994). New Bayesian statistical approaches to estimating and evaluating models of exchange rates determination. In J. S. Meetings, ed., *Proceedings of the ASA Section on Bayesian Statistical Science*. American Statistical Association.

Quintana, J. (1987). *Multivariate Bayesian Forecasting Models*. PhD. Thesis, University of Warwick.

Quintana, J. (1992). Optimal portfolios of forward currency contracts. In

J. Berger, J. Bernardo, A. Dawid, and A. Smith, eds., *Bayesian Statistics IV*. Oxford University Press.

Quintana, J., Chopra, V., and Putnam, B. (1995). Global asset alocation: Stretching returns by shrinking forecasts. In J. S. Meetings, ed., *Proceedings of the ASA Section on Bayesian Statistical Science*. American Statistical Association.

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

Quintana, J. and Putnam, B. (1996). Debating currency markets efficiency using multiple-factor models. In J. S. Meetings, ed., *Proceedings of the ASA Section on Bayesian Statistical Science*. American Statistical Association.

Quintana, J. and 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., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 1197–1208.

Roverato, A. (2000). Cholesky decomposition of a hyper-inverse Wishart matrix. *Biometrika* **87**, 99–112.

Roverato, A. (2002). Hyper-inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics* **29**, 391–411.

Shoker, B., Jarvis, C., Davies, M., Iqbal, M., Sibson, D., and Sloane, J. (2001). Immunodetectable cyclin d1 is associated with oestrogen receptor but not ki67 in normal, cancer and precancerous breast lesions. *British Journal of Cancer* **84**, 1064–1068.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology* **15**, 201–293.

Speed, T. and Kiiveri, H. (1986). Gaussian Markov distributions over finite graphs. *The Annals of Statistics* **14**, 138–150.

Stevens, G. (1998). On the inverse of the covariance matrix in portfolio analysis. *The Journal of Finance* **53**, 1821–1827.

Tukey, P. and Tukey, J. (1981). Graphical display of data sets in 3 or more dimensions. In V. Barnett, ed., *Interpreting Multivariate Data*, 189–257. Wiley, New York.

Uhlig, H. (1994). On singular Wishart and singular multivariate beta distributions. *Annals of Statistics* **22**, 395–405.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In J. M. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, eds., *Bayesian Statistics 7*, 723–732. Oxford University Press.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., J.R., M., and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer using gene expression profiles. *Proceedings of the National Academy of Sciences* **98**, 11462–11467.

West, M. and Harrison, P. (1997). *Bayesian Forecasting and Dynamic Models*. Springer, New York, USA.

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

Wong, F., Carter, C., and Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika* **90**, 809–830.

Wykoff, C., Beasley, N., Watson, P., Campo, L., Chia, S., English, R., Pastorek, J., Sly, W., Ratcliffe, P., and Harris, A. (2001). Expression of the hypoxia-inducible and tumor-associated carbonic anhydrases in ductal carcinoma in situ of the breast. *American Journal of Pathology* **158**, 1011–1019.

Yamashita, H., Nishio, M., Toyama, T., Sugiura, H., Zhang, Z., Kobayashi, S., and Iwase, H. (2004). Coexistence of HER2 over-expression and p53 protein accumulation is a strong prognostic molecular marker in breast cancer. *Breast Cancer Research* **6**, 24–30.

Zelinski, D., Zantek, N., Walker-Daniels, J., Peters, M., Taparowsky, E., and

Kinch, M. (2002). Estrogen and myc negatively regulate expression of the epha2 tyrosine kinase. *Journal of Cellular Biochemistry* **85**, 714–720.

Zellner, A. (1986). *On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions*, chap. 15, 0–0. Amsterdam: Elsevier Science.

Biography

Carlos Marinho Carvalho was born on July 29, 1978 in Petropolis, Rio de Janeiro (Brazil). He received his bachelor degree in economics on December 16, 1999 from IBMEC Business School in Rio de Janeiro (Brazil). He received a M.S. in statistics from the Federal University of Rio de Janeiro on April 1, 2002 and another M.S. in statistics from Duke University in Durham, North Carolina on May 1, 2005. Before coming to Duke, Carlos was an Assistant Professor of Statistics and Econometrics at IBMEC Business School, Rio de Janeiro. He has co-authored the following articles:

1. Carvalho,C.M., Wang, Q., Lucas, J., Chang, J., Nevins, J.R. and West, M. (2005). High-dimensional sparse factor models and latent factor regression. *ISDS Discussion Paper 05-15. Submitted for publication.*

2. Carvalho, C.M., Massam, H. and West, M. (2005). Simulation of hyper-inverse Wishart distributions in graphical models. *ISDS Discussion Paper 05-03. Submitted for publication.*

3. 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 pg388–400.*

4. Lopes, H., Carvalho,C.M. (2005). Factor stochastic volatility with time-varying loadings and Markov switching regimes. *Submitted for publication.*

5. Carvalho, C.M. (2004). Simulation-based sequential analysis of Markov switching stochastic volatility models. *Submitted for publication.*